MaaS

API Reference

Issue 01

Date 2026-07-03

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.

Huawei Cloud Computing Technologies Co., Ltd.

Address: Huawei Cloud Data Center Jiaoxinggong Road
Qianzhong Avenue
Gui'an New District
Gui Zhou 550029
People's Republic of China

Website: https://www.huaweicloud.com/intl/en-us/

Contents

1 Before You Start

Before calling MaaS APIs, ensure that you are familiar with MaaS concepts.

MaaS supports Representational State Transfer (REST) APIs, allowing you to call APIs using HTTPS. For details about API calling, see **Calling APIs**.

Endpoints

An endpoint is the request address for calling an API. Endpoints vary depending on services and regions. For the endpoints of all services, see **Regions and Endpoints**.

Constraints

For more constraints, see API description.

Terms and Definitions

- Account

  An account is created upon successful registration with the cloud platform. The account has full access permissions for all of its cloud services and resources. It can be used to reset user passwords and grant user permissions. The account is a payment entity and should not be used directly to perform routine management. For security purposes, create IAM users and grant them permissions for routine management.

- IAM user

  An IAM user is created by an account in IAM to use cloud services. Each IAM user has its own identity credentials (password and access keys).

  An IAM user can view the account ID and IAM user ID on the **My Credentials** page of the console. The account, username, and password will be required for API authentication.

- Region

  A region is a geographic area in which cloud resources are deployed. Availability zones (AZs) in the same region can communicate with each other over an intranet, while AZs in different regions are isolated from each other. Deploying cloud resources across multiple regions helps meet specific user needs and ensures compliance with local regulations.

- AZ

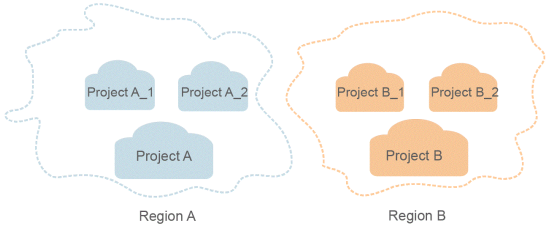
  An AZ contains one or more physical data centers. Each AZ has independent cooling, fire extinguishing, moisture-proof, and electricity facilities. An AZ's compute, network, storage, and other resources are logically divided into multiple clusters. AZs within a region are interconnected by optical fibers for high-availability networking.

- Project

  Projects group and isolate resources (including compute, storage, and network resources) across physical regions. A default project is provided for each the cloud region, and subprojects can be created under each default project. Users can be granted permissions to access all resources in a specific project. If you need more refined access control, create subprojects under a default project and create resources in subprojects. Then you can assign users the permissions required to access only the resources in the specific subprojects.

  To view a project ID, go to the **My Credentials** page.

  Figure 1-1 Project isolation model



2 API Overview

This section describes the APIs supported by MaaS.

Querying ModelArts Real-Time Services

API	Description
Obtaining Region Information	Used to obtain the region information of ModelArts real-time services. When **source** is set to **custom_from_modelarts_v2** during the creation of a custom endpoint, the **region** parameter in the request is obtained.
Obtaining Workspace Information	Used to obtain the workspace information of ModelArts real-time services. When **source** is set to **custom_from_modelarts_v2** during the creation of a custom endpoint, the **infer_service_id** is obtained.
Obtaining Inference Service Information	Used to obtain the workspace information of ModelArts real-time services. When **source** is set to **custom_from_modelarts_v2** during the creation of a custom endpoint, the inference service information is obtained.
Obtaining the Latest Content Guard Disclaimer	Used to obtain the content guard disclaimer of MaaS real-time inference. If **moderation** is set to **false** when you create an endpoint, you must sign the disclaimer of the latest version.

MaaS Call Statistics

API	Description
Obtaining Total Statistics	Used to query the aggregated call data of real-time inference services, including: total calls, total failed calls, total tokens, input tokens, output tokens, etc. Data is retained for 30 days only.

API	Description
Obtaining the Service Statistics List	Used to retrieve statistics across three types of services: subscribed built-in services, created endpoints, and deployed "My Services." It displays metrics for each service, such as call count, failure rate, total tokens, input tokens, output tokens, and E2E latency. Data is retained for 30 days only.
Obtaining Time-based Service Metric Statistics	Used to retrieve granular metric data for a service. This allows viewing chronological trends for metrics such as call count, failure rate, token volume, input token size, output token size, E2E latency, TPM, RPM, QPS, and average generation time. Data is retained for 30 days only.
Obtaining Service Error Details	Used to retrieve detailed error data for services to review failure information, such as error codes, occurrences, and error messages. Data is retained for 30 days only.
Obtaining the Service List	Used to retrieve the service name corresponding to a specific service ID.
Querying the IP Address List	Used to retrieve IP addresses.
Querying Resource Monitoring Metric Details	Used to query the resource monitoring metrics for the "My Services" category within the MaaS real-time inference module. Data is retained for 30 days only.
Querying the Calling Data of a Service Version	Used to query all versions of a service and their corresponding monitoring metric data. Data is retained for 30 days only.
Obtaining the Metrics Supported by Different Model Types	Used to retrieve the list of supported metrics corresponding to a specific service model type.
Obtaining Time-based Service Error Code Statistics	Used to display the chronological distribution of service error code statistics. Data is retained for 30 days only.

3 Calling APIs

3.1 Making an API Request

This section describes the structure of a REST API request, and uses the IAM API for **Obtaining a User Token** as an example to demonstrate how to call an API. The obtained token can then be used to authenticate the calling of other APIs.

Request URI

A request URI is in the following format:

{URI-scheme}://{Endpoint}/{resource-path}?{query-string}

Table 3-1 Request URI

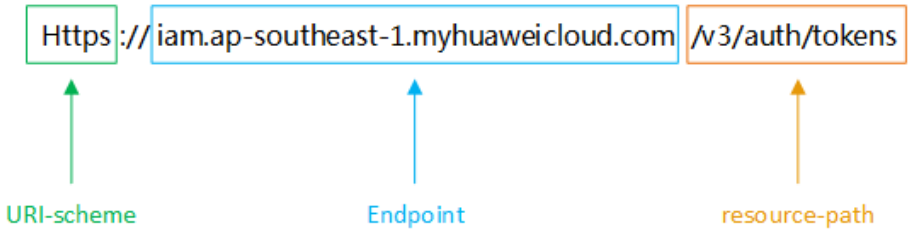
Parameter	Description
URI-scheme	Protocol used to transmit requests. All APIs use HTTPS.
Endpoint	Domain name or IP address of the server for the REST service endpoint. The endpoint varies depending on services in different regions. It can be obtained in **Endpoints**. For example, the endpoint of IAM in the **CN-Hong Kong** region is **iam.ap-southeast-1.myhuaweicloud.com**. **NOTE** Use **iam-endpoint** to obtain a token (**iam-apigateway-proxy.{region_id}.{external_domain_name}**). Configure the domain name in the format of "{float_ip} iam-apigateway-proxy.{region_id}.{external_domain_name}" in the hosts file on the local PC. For details, see **Obtaining a User Token**.
resource-path	Access path of an API for performing a specified operation. Obtain the path from the URI of an API. For example, the **resource-path** of the API used to obtain a user token is **/v3/auth/tokens**.

Parameter	Description
query-string	Query parameter, which is optional. Ensure that a question mark (?) is included before a query parameter that is in the format of "Parameter name=Parameter value". For example, **? limit=10** indicates that a maximum of 10 data records will be displayed.

For example, to obtain a token in the **CN-Hong Kong** region, obtain the endpoint of IAM (**iam.ap-southeast-1.myhuaweicloud.com**) for this region and the **resource-path** (**/v3/auth/tokens**) in the URI of the API used to **obtain a user token**. Then, construct the URI as follows:

```
https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens
```

Figure 3-1 Example URI



NOTE

To simplify the URI display in this document, each API is provided only with a **resource-path** and a request method. The **URI-scheme** of all APIs is **HTTPS**, and the endpoints of all APIs in the same region are identical.

Request Methods

HTTP defines the following request methods that can be used to send a request to the server.

Table 3-2 HTTP methods

Method	Description
GET	Requests the server to return specified resources.
PUT	Requests the server to update specified resources.
POST	Requests the server to add resources or perform special operations.
DELETE	Requests the server to delete specified resources, for example, an object.
HEAD	Requests the server to return the response header.

Method	Description
PATCH	Requests the server to update partial content of a specified resource. If the resource does not exist, a new resource will be created.

For example, in the URI of the API for **obtaining a user token**, the request method is POST. The request is as follows:

```
POST https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens
```

Request Headers

You can also add additional header fields to a request, such as the fields required by a specified URI or HTTP method. For example, to request for the authentication information, add **Content-Type**, which specifies the request body type.

Table 3-3 describes common request header fields.

Table 3-3 Common request headers

Parameter	Description	Mandatory	Example Value
Content-type	Request body type or format. The default value is **application/json**.	Yes	application/json
Content-Length	Length of the request body. The unit is byte.	Mandatory for POST and PUT requests but must be left blank for GET requests	3495
X-Project-Id	Project ID. This parameter is used to obtain the token for each project.	No	e9993fc787d94b6c886cbaa340f9c0f4
X-Auth-Token	The user token is a response to the API used to **obtain a user token**. This API is the only one that does not require authentication.	Mandatory for token-based authentication	-

Parameter	Description	Mandatory	Example Value
X-Sdk-Date	Time when the request is sent. The time is in *YYYYMMDD'T'HHMMSS'Z'* format. The value is the current Greenwich Mean Time (GMT) time of the system.	Mandatory for AK/SK-based authentication, optional for PKI token-based authentication	20190307T101459Z
Authorization	Authentication information. The value is obtained from the request signature result and is required when the AK/SK are used to encrypt the signature. Type: string Default value: none	Mandatory for AK/SK-based authentication	SDK-HMAC-SHA256 Credential=ZIRRKMTWPTQFQI1WKNKB/20150907//ec2/sdk_request, SignedHeaders=content-type;host;x-sdk-date, Signature=55741b610f3c9fa3ae40b5a8021ebf7ebc2a28a603fc62d25cb3bfe6608e1994
Host	Information about the requested server. The value can be obtained from the URL of the service API. The value is in the format of *hostname:port*. If the port number is not specified, the default port is used. The default port number for **https** is **443**.	Mandatory for AK/SK-based authentication	code.test.com or code.test.com:443

NOTE

In addition to supporting token-based authentication, APIs support authentication using AK/SK. During AK/SK authentication, an SDK is used to sign a request, and the **Authorization** (signature authentication) and **X-Sdk-Date** (time when a request is sent) headers are automatically added to the request. For more information about AK/SK-based authentication, see **AK/SK Signing and Authentication Guide**.

The API used to **obtain a user token** does not require authentication. Therefore, only the **Content-Type** field needs to be added to requests for calling the API. An example of such requests is as follows:

```
POST https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens
Content-Type: application/json
```


Request Body

The body of a request is often sent in a structured format as specified in the **Content-Type** header field. The request body transfers content except the request header. If the request body contains Chinese characters, these characters must be encoded in UTF-8.

The request body varies depending on APIs. Some APIs do not require the request body, such as the APIs requested using the GET or DELETE method.

If an API is used to **obtain a user token**, the request parameters and parameter description can be obtained from the API request. The request with the request body added is as follows. Replace the italic fields in bold with the actual values. *user_name* indicates the username, *domain_name* indicates the account name, *user_password* indicates the login password, and **ap-southeast-1** indicates the project name. For details about how to obtain the project name, see **Obtaining a Username and ID**, **Obtaining an Account Name and Account ID**, and **Obtaining a Username**.

NOTE

The **scope** parameter specifies where a token takes effect. In the example, the token takes effect only for the resources in a specified project. MaaS uses a region-specific endpoint to call this API. Set **scope** to **project**. You can set **scope** to an account or a project under an account. For details, see **Obtaining a User Token**.

```
POST https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens
Content-Type:application/json
{
  "auth": {
    "identity": {
      "methods": ["password"],
      "password": {
        "user": {
          "name": "user_name",
          "password": "user_password",
          "domain": {
            "name": "domain_name"
          }
        }
      }
    },
    "scope": {
      "project": {
        "name": "ap-southeast-1"
      }
    }
  }
}
```

After all data required for the API request is available, send the request to call the API through **curl**, **Postman**, or coding. In the response to the API used to **obtain a user token**, **x-subject-token** is the desired user token. This token can then be used to authenticate the calling of other APIs.

3.2 Authentication

API call requests can be authenticated using tokens.

Token-based Authentication

NOTE

The validity period of a token is 24 hours. When using a token for authentication, cache it to prevent frequently calling the IAM API used to obtain a user token.

A token specifies certain permissions in a computer system. During API authentication using a token, the token is added to requests to get permissions for calling the API.

When calling the API to **obtain a user token**, you must set **auth.scope** in the request body to **project**.

In **Making an API Request**, the process of calling the API used to obtain a user token is described.

```
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "user_name",
                    "password": "user_password",
                    "domain": {
                        "name": "domain_name"
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "project_name"
            }
        }
    }
}
```

After a token is obtained, the **X-Auth-Token** header must be added to requests to specify the token when calling other APIs. For example, if the token is **ABCDEFJ....**, add **X-Auth-Token: ABCDEFJ....** to a request as follows:

```
GET https://modelarts.ap-southeast-1.myhuaweicloud.com/v1/{project_id}/services
Content-Type: application/json
X-Auth-Token: ABCDEFJ....
```

3.3 Responses

After sending a request, you will receive a response, including the status code, response header, and response body.

Status Code

A status code is a group of digits, ranging from 1*xx* to 5*xx*. It indicates the status of a request. For more information, see **Status Codes**.

If status code **201** is returned for calling the API used to **obtain a user token**, the request is successful.

Table 3-4 Returned status code

Return Value	Description
201	Creation succeeded.
400	Invalid parameters.
401	Authentication failed.
403	No operation permission.
404	Requested resource cannot found.

Response Headers

Similar to a request, a response also has a header, for example, **Content-type**.

Header	Description	Mandatory
Content-Type	Media type of the response body sent to a recipient. Type: string. Default value: **application/json; charset=UTF-8**	Yes
X-request-id	Request ID for task tracing. Type: string. **request_id-timestamp-hostname** (**request_id** is the UUID generated on the server, **timestamp** indicates the current timestamp, and **hostname** is the name of the server that processes the current API.) Default value: none	No
X-ratelimit	Total number of throttling requests. Type: integer. Default value: none.	No
X-ratelimit-used	Number of remaining requests Type: integer. Default value: none.	No
X-ratelimit-window	Throttling unit. Type: string. The unit is minute, hour, or day. Default value: hour.	No

Figure 1 shows the response header for the API used to **obtain a user token**.

x-subject-token is the desired user token. This token can then be used to authenticate the calling of other APIs.

Figure 3-2 Header fields of the response to the request for obtaining a user token

Response Body

The body of a response is often returned in structured format as specified in the **Content-Type** header field. The response body transfers content except the response header.

For the API used to **obtain a user token**, the response body is as follows: The following shows part of the response body for the API to obtain a user token.

```
{
    "token": {
        "expires_at": "2019-02-13T06:52:13.855000Z",
        "methods": [
            "password"
        ],
        "catalog": [
            {
                "endpoints": [
                    {
                        "region_id": "ap-southeast-1",
.....
```

If an error occurs during API calling, an error code and error message will be displayed. The following shows the response body in the case of an error:

```
{
    "error_msg": "The format of message is error",
    "error_code": "AS.0001"
}
```

In the error response body, **error_code** is an error code, and **error_msg** provides information about the error. For more details, see **Error Codes**.

4 Querying ModelArts Real-Time Services

4.1 Obtaining Region Information

Description

This API is used to obtain the region information of ModelArts real-time services. When **source** is set to **custom_from_modelarts_v2** during the creation of a custom endpoint, the **region** parameter in the request is obtained.

Constraints

This function is only supported in CN-Hong Kong.

URI

GET /v1/{project_id}/maas/services/custom-endpoint/regions

Table 4-1 URI parameters

Parameter	Mandatory	Type	Description
project-id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A. **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A.

Request Parameters

Table 4-2 Request header parameters

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. The token can be obtained by calling the IAM API used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details about how to obtain the value, see **Authentication**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.
Content-Type	Yes	String	**Definition**: Type of the message body. The value is fixed to **application/json**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.

Response Parameters

Table 4-3 Response body parameters

Parameter	Type	Description
region_items	Array[RegionInfo]	**Definition**: All region information of ModelArts real-time services. **Range**: N/A.

Table 4-4 RegionInfo

Parameter	Type	Description
region_id	String	**Definition**: Region of ModelArts real-time services. **Range**: N/A.
locales	LocalInfo	**Definition**: Endpoint name. **Range**: N/A.

Table 4-5 LocalInfo

Parameter	Type	Description
en_us	String	**Definition**: English name. **Range**: N/A.

Table 4-6 Error response parameters

Parameter	Type	Description
error_msg	String	**Definition**: Error description. **Range**: N/A.
error_code	String	**Definition**: Error code, indicating the error type. **Range**: N/A.

Request Example

```
GET
/v1/{project_id}/maas/services/custom-endpoint/regions
```

Response Example

- Success response. Status code: 200.

```
{
    "region_items": [
        {
            "region_id": "cn-southwest-2",
            "locales": {
                "en_us": "cn-southwest-2",
            }
        },
        {
            "region_id": "cn-north-12",
            "locales": {
                "en_us": "cn-north-12",
            }
        },
        {
            "region_id": "cn-east-4",
            "locales": {
                "en_us": "cn-east-4",
            }
        },
        {
            "region_id": "cn-north-9",
            "locales": {
                "en_us": "cn-north-9",
            }
        }
    ]
}
```


- Error response. Status code: 400.

```
{
"error_msg": "Invalid token.",
"error_code": "ModelArts.0104"
}
```

Status Codes

For details, see **Status Codes**.

Error Codes

For details, see **Error Codes**.

4.2 Obtaining Workspace Information

Description

This API is used to obtain the workspace information of ModelArts real-time services. When **source** is set to **custom_from_modelarts_v2** during the creation of a custom endpoint, the **infer_service_id** is obtained.

Constraints

This function is only supported in CN-Hong Kong.

URI

GET /v1/{project_id}/maas/services/workspace/{region_id}

Table 4-7 URI parameters

Parameter	Mandatory	Type	Description
project-id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.

Parameter	Mandatory	Type	Description
region_id	Yes	String	**Definition**: Region information of ModelArts real-time services. For details about how to obtain the value, see **Obtaining Region Information**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.

Request Parameters

Table 4-8 Request header parameters

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. The token can be obtained by calling the IAM API used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.
Content-Type	Yes	String	**Definition**: Type of the message body. The value is fixed to **application/json**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.

Response Parameters

Table 4-9 Response body parameters

Parameter	Type	Description
total_count	Integer	**Definition**: Total number of workspaces that can be queried. **Range**: N/A.
count	Integer	**Definition**: Number of workspaces returned in this query. **Range**: N/A.
workspaces	Array[WorkspaceInfo]	**Definition**: Workspace information. **Range**: N/A.

Table 4-10 WorkspaceInfo

Parameter	Type	Description
id	String	**Definition**: Workspace ID. **Range**: N/A.
name	String	**Definition**: Workspace name. **Range**: N/A.
description	String	**Definition**: Workspace description. **Range**: N/A.
owner	String	**Definition**: Workspace owner. **Range**: N/A.
create_time	String	**Definition**: Creation time. **Range**: N/A.
update_time	String	**Definition**: Update time. **Range**: N/A.
auth_type	String	**Definition**: Authentication type. **Range**: N/A.

Table 4-11 Error response parameters

Parameter	Type	Description
error_msg	String	**Definition**: Error description. **Range**: N/A.
error_code	String	**Definition**: Error code, indicating the error type. **Range**: N/A.

Response Example

- Success response. Status code: 200.

```
{
   "total_count": 1,
   "count": 1,
   "workspaces": [
      {
         "id": "0",
         "name": "default",
         "description": "",
         "owner": "******",
         "create_time": 1680746333000,
         "update_time": 1764128642000,
         "auth_type": "public"
      }
   ]
}
```

- Error response. Status code: 400.

```
{
"error_msg": "Invalid token.",
"error_code": "ModelArts.0104"
}
```

Status Codes

For details, see **Status Codes**.

Error Codes

For details, see **Error Codes**.

4.3 Obtaining Inference Service Information

Description

This API is used to obtain the information of ModelArts real-time inference services. When **source** is set to **custom_from_modelarts_v2** during the creation of a custom endpoint, the inference service information is obtained.

Constraints

This function is only supported in CN-Hong Kong.

URI

GET /v1/{project_id}/maas/services/custom-endpoint/services/{region_id}?workspace_id={workspace_id}

Table 4-12 URI parameters

Parameter	Mandatory	Type	Description
project-id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.
region_id	Yes	String	**Definition**: Region information of ModelArts real-time services. For details about how to obtain the value, see **Obtaining Region Information**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.
workspace_id	No	String	**Definition**: ID of the workspace whose resources are to be queried. If no value is specified, the default workspace is queried. For details about how to obtain the value, see **Obtaining Workspace Information**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.

Request Parameters

Table 4-13 Request header parameters

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. The token can be obtained by calling the IAM API used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.
Content-Type	Yes	String	**Definition**: Type of the message body. The value is fixed to **application/json**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.

Response Parameters

Table 4-14 Response body parameters

Parameter	Type	Description
data	Array[InferServerInfo]	**Definition**: Inference service information. **Range**: N/A.
pages	Integer	**Definition**: Total number of pages. **Range**: N/A.
total	Integer	**Definition**: Total number of inference information records. **Range**: N/A.

Table 4-15 InferServerInfo

Parameter	Type	Description
id	String	**Definition**: Inference service ID. **Range**: N/A.
name	String	**Definition**: Inference service name. **Range**: N/A.
status	String	**Definition**: Inference service status. **Range**: N/A.
version	String	**Definition**: Version. **Range**: N/A.
version_count	String	**Definition**: Total number of versions. **Range**: N/A.
description	String	**Definition**: Inference service description. **Range**: N/A.
type	String	**Definition**: Inference service type. **Range**: N/A.
deploy_type	String	**Definition**: Deployment type. **Range**: N/A.
user_name	String	**Definition**: Username. **Range**: N/A.
workspace_id	String	**Definition**: Workspace. **Range**: N/A.
create_at	String	**Definition**: Creation time. **Range**: N/A.
update_at	String	**Definition**: Update time. **Range**: N/A.
auth_type	String	**Definition**: Authentication type. **Range**: N/A.

Table 4-16 Error response parameters

Parameter	Type	Description
error_msg	String	**Definition**: Error description. **Range**: N/A.
error_code	String	**Definition**: Error code, indicating the error type. **Range**: N/A.

Request Example

```
GET
/v1/{project_id}/maas/services/custom-endpoint/services/{region_id}?workspace_id={workspace_id}
```

Response Example

- Success response. Status code: 200.

```
{
   "data": [
      {
         "id": "add6b9f8-7e97-4f1c-8816-************",
         "name": "dpsk-v3_2-*****",
         "status": "RUNNING",
         "version": "0.0.9",
         "version_count": 10,
         "description": "DeepSeek-V3.2-EXP",
         "type": "REAL_TIME",
         "deploy_type": "MULTI",
         "user_name": "******",
         "workspace_id": "0",
         "create_at": 1760538261106,
         "update_at": 1765711080510,
         "auth_type": "NONE"
      }
   ],
   "pages": 1,
   "total": 1,
}
```

- Error response. Status code: 400.

```
{
"error_msg": "Invalid token.",
"error_code": "ModelArts.0104"
}
```

Status Codes

For details, see **Status Codes**.

Error Codes

For details, see **Error Codes**.

4.4 Obtaining the Latest Content Guard Disclaimer

Description

This API is used to obtain the content guard disclaimer of MaaS real-time inference. If **moderation** is set to **false** when you create an endpoint, you must sign the disclaimer of the latest version.

Constraints

This function is only supported in CN-Hong Kong.

URI

GET /v1/{project_id}/maas/compliance/agreements

Table 4-17 URI parameters

Parameter	Mandatory	Type	Description
project-id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.

Request Parameters

Table 4-18 Request header parameters

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. The token can be obtained by calling the IAM API used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.

Parameter	Mandatory	Type	Description
Content-Type	Yes	String	**Definition**: Type of the message body. The value is fixed to **application/json**. **Constraints**: N/A. **Range**: N/A. **Default Value**: N/A.

Table 4-19 Query parameters

Parameter	Mandatory	Type	Description
lang	Yes	String	**Definition**: Disclaimer language. Obtain the disclaimer in the required language. **Constraints**: N/A. **Range**: • **zh**: Chinese • **en**: English **Default Value**: N/A.

Response Parameters

Table 4-20 Response body parameters

Parameter	Type	Description
id	String	**Definition**: ID. **Range**: N/A.
agreement_id	String	**Definition**: ID of the disclaimer, which corresponds to the **agreement_id** parameter when you create or edit an endpoint. **Range**: N/A.
agreement_data	String	**Definition**: Disclaimer content. **Range**: N/A.
open_notice	String	**Definition**: Enables content guard notifications. **Range**: N/A.

Parameter	Type	Description
close_notice	String	**Definition**: Disables content guard notifications. **Range**: N/A.

Table 4-21 Error response parameters

Parameter	Type	Description
error_msg	String	**Definition**: Error description. **Range**: N/A.
error_code	String	**Definition**: Error code, indicating the error type. **Range**: N/A.

Request Example

```
GET
/v1/{project_id}/maas/compliance/agreements?lang={lang}
```

Response Example

- Success response. Status code: 200.

```
{
    "agreement_id": "4dd6b9f8-7e97-4f1c-8816-**********",
    "agreement_data": "Create a disclaimer",
    "open_notice": "Huawei Cloud provides no warranties or guarantees regarding third-party models' outputs. It does not ensure their compliance, security, authenticity, accuracy, or completeness. These outputs do not represent Huawei Cloud's views.",
    "close_notice": "The company/individual hereby requests to disable content moderation due to business needs. The company/individual fully understands that Huawei Cloud will disable content moderation associated with the purchased MaaS platform. After the feature is disabled, Huawei Cloud will no longer provide any content moderation capability. |After Huawei Cloud disables content moderation, the company will independently conduct content moderation for all input and output content involved in the MaaS platform. The company will ensure that the model's input, output, and related usage behaviors comply with applicable laws and regulations, including but not limited to the *Cybersecurity Law of the People's Republic of China* , the *Data Security Law of the People's Republic of China* , the *Interim Measures to Regulate Generative AI Services* , and the *Provisions on the Administration of Deep Synthesis in Internet Information Services*. |The company will assume all responsibilities and adverse consequences arising from the disabling of content moderation, including but not limited to losses suffered by the company or its end users, as well as penalties imposed by regulatory authorities. If Huawei Cloud incurs any losses as a result, the company will be liable for compensation. |If the company's input or output content on the MaaS platform fails to comply with relevant laws and regulations, Huawei Cloud has the right to suspend or terminate the provision of services.",
    "id": 1
}
```

- Error response. Status code: 400.

```
{
"error_msg": "Invalid token.",
"error_code": "ModelArts.0104"
}
```


Status Codes

For details, see **Status Codes**.

Error Codes

For details, see **Error Codes**.

5 MaaS Call Statistics

5.1 Obtaining Total Statistics

Function

Used to query the aggregated call data of real-time inference services, including: total calls, total failed calls, total tokens, input tokens, output tokens, etc. Data is retained for 30 days only.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 1,000 within one minute.
- User request limit: The number of requests for this API from a single user cannot exceed 200 within one minute.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 60 seconds and try again.

URI

POST /v1/{project_id}/maas/monitoring/show-statistics

Table 5-1 URI parameter

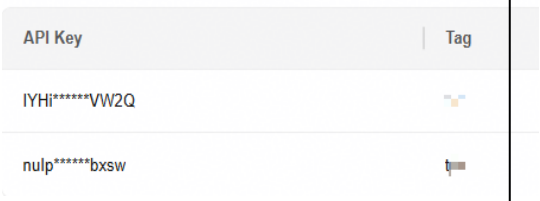
Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A

Request Parameters

Table 5-2 Request header parameter

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Table 5-3 Request body parameters

Parameter	Mandatory	Type	Description
service_type	Yes	Integer	**Definition**: Service type. **Constraints**: N/A **Range**: • **1**: User services. Deploy the model service in the **My Services** tab. For details, see **Deploying a Model Service**. • **2**: Built-in services. Deploy the model service in the **Built-in Services** tab. For details, see **Subscribing to a Built-in Service**. **Default Value**: N/A
api_keys	No	Array of strings	**Definition**: API key tag list, which is used for filtering. MaaS services support **API key calls**. Go to the **API key management** page to get the API key tag. The **Tag** field in the API key list shows the API key tag.  **Constraints**: N/A **Range**: • Example for obtaining the online experience call data: **api_keys: [""]**. • When obtaining the call data of some API keys, transfer the tags of the corresponding API keys. Example: **api_keys: ["test01","test02"]**. • Example for obtaining the call data of online experience and some API keys: **api_keys: ["test01", "test02", ""]**. • When obtaining all call data, do not use this parameter. **Default Value**: N/A

Parameter	Mandatory	Type	Description
ips	No	Array of strings	**Definition**: IP address list, which indicates the source IP addresses of the clients that have been called. If this parameter is not specified, all IP addresses are queried. To query the IP address, you can **call the API for obtaining the IP address list**. **Constraints**: The value must be in the IP address format. **Range**: N/A **Default Value**: N/A
start_time	Yes	Long	**Definition**: Timestamp of the start time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0** and no greater than the value of **end_time**. **Default Value**: N/A
end_time	Yes	Long	**Definition**: Timestamp of the end time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0**. **Default Value**: N/A
timezone	No	String	**Definition**: Time zone. **Constraints**: N/A **Range**: The value must comply with the IANA time zone specifications, such as Asia/Shanghai and UTC. **Default Value**: **Asia/Shanghai (GMT +8)**
infer_type	Yes	String	**Definition**: Service inference type. **Range**: • **real_time**: real-time inference • **batch**: batch inference (Batch inference is under restricted use. To use it, submit a service ticket.) **Constraints**: N/A **Default Value**: N/A

Parameter	Mandatory	Type	Description
model_type	No	string	**Definition**: Model type. **Range**: • **Text Generation**: text generation model • **Video Generation**: video generation model • **Image Generation**: image generation model • **Vector Model**: text vectorization • **Embedding**: embedding model • **Image Understanding**: image understanding model • **Rerank**: reranking model **Constraints**: N/A **Default Value**: **Text Generation**

Response Parameters

Status code: 200

Table 5-4 Response body parameters

Parameter	Type	Description
total_request_count	Integer	**Definition**: Total number of calls. **Range**: N/A
total_error_count	Integer	**Definition**: Total number of failed calls. **Range**: N/A
total_token	Double	**Definition**: Total number of called tokens. If batch inference is used, this parameter indicates the total number of inference tokens. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
total_prompt_token	Double	**Definition**: Total number of input tokens. **Range**: N/A

Parameter	Type	Description
total_completion_token	Double	**Definition**: Total number of output tokens. **Range**: N/A
total_completion_tasks	Integer	**Definition**: Number of completed batch inference tasks. (This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket.) **Range**: N/A
total_infer_count	Integer	**Definition**: Total number of inference times, that is, the total number of service inference times. (This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket.) **Range**: N/A
video_generate_duration	Double	**Definition**: Total duration of the generated video. **Range**: N/A
image_generate_nums	Integer	**Definition**: Total number of generated images. **Range**: N/A

Status code: 400

Table 5-5 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Query the calling statistics of a user-defined endpoint service whose real-time inference time ranges from 1761753600000 to 1761806407404.


```
/v1/{{project_id}}/maas/monitoring/show-statistics

{
  "service_type" : 4,
  "start_time" : 1761753600000,
  "end_time" : 1761806407404,
  "timezone" : "Asia/Shanghai",
  "infer_type" : "real_time"
}
```

Response Example

Status code: 200

Success response

```
{
  "total_request_count" : 202,
  "total_error_count" : 6,
  "total_token" : 78.035,
  "total_prompt_token" : 70.265,
  "total_completion_token" : 7.77,
  "total_completion_tasks" : 0,
  "total_infer_count" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}
```

Status code: 400

Failure response

```
{
  "error_code" : "ModelArts.0104",
  "error_msg" : "Invalid parameter. Issue: The end time cannot be earlier than the start time."
}
```

Status Codes

Status Code	Description
200	Success response.
400	Failure response.

Error Codes

For details, see **Error Codes**.

5.2 Obtaining the Service Statistics List

Function

This API is used to obtain the provisioned preset services or deployed user services, and display metrics such as service calling times, calling failure rate, total number of called tokens, number of input/output tokens, and E2E latency. Data is retained for 30 days only.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 1,000 within one minute.
- User request limit: The number of requests for this API from a single user cannot exceed 200 within one minute.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 60 seconds and try again.

URI

POST /v1/{project_id}/maas/monitoring/list-service-statistics

Table 5-6 URI parameter

Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A

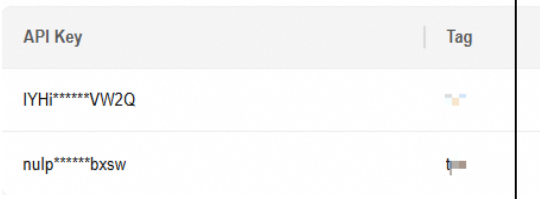
Request Parameters

Table 5-7 Request header parameter

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Table 5-8 Request body parameters

Parameter	Mandatory	Type	Description
service_names	No	Array of strings	**Definition**: Service name list, which can be used to filter services by name. If this parameter is not specified, all services are queried. Fuzzy match of service names is supported. **Constraints**: N/A **Range**: The service list contains provisioned preset services and deployed user services. **Default Value**: N/A
service_type	Yes	Integer	**Definition**: Service type. **Constraints**: N/A **Range**: • **1**: User services. Deploy the model service in the **My Services** tab. For details, see **Deploying a Model Service**. • **2**: Built-in services. Deploy the model service in the **Built-in Services** tab. For details, see **Subscribing to a Built-in Service**. **Default Value**: N/A

Parameter	Mandatory	Type	Description
api_keys	No	Array of strings	**Definition**: API key tag list, which is used for filtering. MaaS services support **API key calls**. Go to the **API key management** page to get the API key tag. The **Tag** field in the API key list shows the API key tag.  **Constraints**: N/A **Range**: • Example for obtaining the online experience call data: **api_keys: [""]**. • When obtaining the call data of some API keys, transfer the tags of the corresponding API keys. Example: **api_keys: ["test01","test02"]**. • Example for obtaining the call data of online experience and some API keys: **api_keys: ["test01", "test02", ""]**. • When obtaining all call data, do not use this parameter. **Default Value**: N/A
ips	No	Array of strings	**Definition**: IP address list, which indicates the source IP addresses of the clients that have been called. If this parameter is not specified, all IP addresses are queried. To query the IP address, you can **call the API for obtaining the IP address list**. **Constraints**: The value must be in the IP address format. **Range**: N/A **Default Value**: N/A

Parameter	Mandatory	Type	Description
start_time	Yes	Long	**Definition**: Timestamp of the start time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0** and no greater than the value of **end_time**. **Default Value**: N/A
end_time	Yes	Long	**Definition**: Timestamp of the end time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0**. **Default Value**: N/A
limit	No	Integer	**Definition**: Number of records on each page, that is, maximum number of returned records. If this parameter is set to **0**, all data is returned without pagination. **Constraints**: N/A **Range**: The value must be greater than or equal to 0. **Default Value**: **0**
offset	No	Integer	**Definition**: Pagination offset, which indicates the number of records to be skipped. **Constraints**: N/A **Range**: The value must be greater than or equal to 0. **Default Value**: **0**
timezone	No	String	**Definition**: Time zone. **Constraints**: N/A **Range**: The value must comply with the IANA time zone specifications, such as Asia/Shanghai and UTC. **Default Value**: **Asia/Shanghai (GMT +8)**

Parameter	Mandatory	Type	Description
infer_type	Yes	String	**Definition**: Service inference type. **Range**: • **real_time**: real-time inference • **batch**: batch inference (Batch inference is under restricted use. To use it, submit a service ticket.) **Constraints**: N/A **Default Value**: N/A
model_type	No	string	**Definition**: Model type. **Range**: • **Text Generation**: text generation model • **Video Generation**: video generation model • **Image Generation**: image generation model • **Vector Model**: text vectorization • **Embedding**: embedding model • **Image Understanding**: image understanding model • **Rerank**: reranking model **Constraints**: N/A **Default Value**: **Text Generation**

Response Parameters

Status code: 200

Table 5-9 Response body parameters

Parameter	Type	Description
total	Integer	**Definition**: Total number of queried data records. **Range**: N/A
count	Integer	**Definition**: Maximum number of data records returned on each page. **Range**: If pagination is used, the value is that of the input parameter **limit**. If pagination is not used, the value is the total number of queried data records.

Parameter	Type	Description
items	Array of **ServiceStatItem** objects	**Definition**: Service information list. **Range**: The service list displays only the provisioned preset services, created custom endpoints, and deployed user-defined services.

Table 5-10 ServiceStatItem

Parameter	Type	Description
service_id	String	**Definition**: Service ID. **Range**: N/A
service_name	String	**Definition**: Service name. **Range**: N/A
request_count	Integer	**Definition**: Number of calls. **Range**: N/A
error_count	Integer	**Definition**: Number of failed calls. If batch inference is used, this parameter indicates the number of failed inference times. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
error_rate	Double	**Definition**: Call failure rate. If batch inference is used, this parameter indicates the inference failure rate. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: **0** to **1**
total_token	Double	**Definition**: Total number of called tokens, in thousands. If batch inference is used, this parameter indicates the total number of tokens. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
prompt_token	Double	**Definition**: Number of input tokens, in thousands. **Range**: N/A

Parameter	Type	Description
completion_token	Double	**Definition**: Number of output tokens **Range**: N/A
avg_latency	Double	**Definition**: Average E2E latency (ms) **Range**: N/A
avg_ttft	Double	**Definition**: Average time to first token (TTFT) (ms) **Range**: N/A
avg_tpot	Double	**Definition**: Average time per output token (TPOT) (ms) **Range**: N/A
infer_times	Integer	**Definition**: Total number of inference times. (This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket.) **Range**: N/A
scc_count	Integer	**Definition**: Number of successful calls. If batch inference is used, this parameter indicates the number of successful inference times. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
avg_consume_time	Double	**Definition**: Average task handling duration, in minutes. This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
completion_tasks_count	Integer	**Definition**: Number of completed tasks. This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
cache_token	Double	**Definition**: Number of cache hits (thousand tokens). **Range**: N/A

Parameter	Type	Description
cache_hit_ratio	Double	**Definition**: Cache hit rate, which is number of cache hit tokens/number of input tokens. **Range**: N/A
avg_generation_time	Double	**Definition**: Average generation duration, that is, the average time generating an image or video. Only video and image models are supported. **Range**: N/A
generation_type	String	**Definition**: Model type. **Range**: • **Text Generation**: text generation model • **Video Generation**: video generation model • **Image Generation**: image generation model • **Vector Model**: text vectorization • **Embedding**: embedding model • **Image Understanding**: image understanding model • **Rerank**: reranking model
video_generate_duration	Double	**Definition**: Total duration of the generated video. **Range**: N/A
image_generate_nums	Integer	**Definition**: Total number of generated images. **Range**: N/A

Status code: 400

Table 5-11 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A

Parameter	Type	Description
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Query the calling statistics of a preset service whose real-time inference time ranges from 1770048000000 to 1770103717647. A maximum of 100 records can be displayed on each page.

```
v1/{{project_id}}/maas/monitoring/list-service-statistics

{
  "start_time" : 1770048000000,
  "end_time" : 1770103717647,
  "service_type" : 2,
  "limit" : 100,
  "offset" : 0,
  "service_names" : [ ],
  "infer_type" : "real_time",
  "model_type" : "Text Generation"
}
```

Response Example

Status code: 200

Success response

```
{
  "total" : 6,
  "count" : 100,
  "items" : [ {
    "service_id" : "21699c35-f333-462b-8a5d-66599926c26f",
    "service_name" : "DeepSeek-R1",
    "generation_type" : "Text Generation",
    "request_count" : 32,
    "error_count" : 0,
    "error_rate" : 0,
    "total_token" : 47.872,
    "prompt_token" : 15.104,
    "completion_token" : 32.768,
    "avg_latency" : 33548.47,
    "avg_ttft" : 343.83,
    "avg_tpot" : 32.46,
    "avg_generation_time" : 0,
    "completion_tasks_count" : 0,
    "infer_times" : 0,
    "scc_count" : 0,
    "avg_consume_time" : 0,
    "cache_token" : 0,
    "cache_hit_ratio" : 0,
    "video_generate_duration": 0,
    "image_generate_nums": 0
  }, {
    "service_id" : "3f23f78d-96e3-4146-885a-74fc392ed190",
    "service_name" : "DeepSeek-V3.2",
    "generation_type" : "Text Generation",
    "request_count" : 35,
    "error_count" : 0,
```



```
  "error_rate" : 0,
  "total_token" : 68.11,
  "prompt_token" : 32.27,
  "completion_token" : 35.84,
  "avg_latency" : 35744.14,
  "avg_ttft" : 1139.02,
  "avg_tpot" : 33.73,
  "avg_generation_time" : 0,
  "completion_tasks_count" : 0,
  "infer_times" : 0,
  "scc_count" : 0,
  "avg_consume_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "service_id" : "44e8ee1d-890e-4f8d-9b05-2b2f03a9e514",
  "service_name" : "DeepSeek-V3",
  "generation_type" : "Text Generation",
  "request_count" : 17,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 33.065,
  "prompt_token" : 15.657,
  "completion_token" : 17.408,
  "avg_latency" : 30236.35,
  "avg_ttft" : 328.66,
  "avg_tpot" : 29.24,
  "avg_generation_time" : 0,
  "completion_tasks_count" : 0,
  "infer_times" : 0,
  "scc_count" : 0,
  "avg_consume_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "service_id" : "780caccf-c894-4fd9-8a8e-ba31f1c644fe",
  "service_name" : "Qwen3-32B-32K",
  "generation_type" : "Text Generation",
  "request_count" : 32,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 70.316,
  "prompt_token" : 32.48,
  "completion_token" : 37.836,
  "avg_latency" : 21912.5,
  "avg_ttft" : 349.18,
  "avg_tpot" : 18.2,
  "avg_generation_time" : 0,
  "completion_tasks_count" : 0,
  "infer_times" : 0,
  "scc_count" : 0,
  "avg_consume_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "service_id" : "8f658111-64af-44dd-bd3b-a78ec500bc88",
  "service_name" : "DeepSeek-V3.2-Exp",
  "generation_type" : "Text Generation",
  "request_count" : 32,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 57.17,
  "prompt_token" : 15.088,
```



```
  "completion_token" : 42.082,
  "avg_latency" : 36555.44,
  "avg_ttft" : 0,
  "avg_tpot" : 0,
  "avg_generation_time" : 0,
  "completion_tasks_count" : 0,
  "infer_times" : 0,
  "scc_count" : 0,
  "avg_consume_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
 }, {
  "service_id" : "ecd6ff0d-4634-4a50-bbe5-641c27b26087",
  "service_name" : "DeepSeek-V3.1",
  "generation_type" : "Text Generation",
  "request_count" : 32,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 59.302,
  "prompt_token" : 29.568,
  "completion_token" : 29.734,
  "avg_latency" : 31596.78,
  "avg_ttft" : 504.26,
  "avg_tpot" : 32.93,
  "avg_generation_time" : 0,
  "completion_tasks_count" : 0,
  "infer_times" : 0,
  "scc_count" : 0,
  "avg_consume_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
 } ]
}
```


Status code: 400

Failure response


```
{
  "error_code" : "ModelArts.0104",
  "error_msg" : "Invalid parameter. Issue: The end time cannot be earlier than the start time."
}
```


Status Codes

Status Code	Description
200	Success response
400	Failure response

Error Codes

For details, see **Error Codes**.

5.3 Obtaining Time-based Service Metric Statistics

Function

This API is used to obtain service metric details. This allows you to view chronological trends for metrics such as call count, failure rate, token volume, input token size, output token size, E2E latency, TPM, RPM, QPS, and average generation time. Data is retained for 30 days only.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 80 within 20 seconds.
- User request limit: The number of requests for this API from a single user cannot exceed 1 within 20 seconds.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 20 seconds and try again.

URI

POST /v1/{project_id}/maas/monitoring/{service_id}/show-detail-chart

Table 5-12 URI parameters

Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A

Parameter	Mandatory	Type	Description
service_id	Yes	String	**Definition**: Service IDs to be queried. Services are filtered based on the input service ID list. If this parameter is not specified, all service names corresponding to the IDs are returned. You can obtain the service ID from the response body during service creation, or **call the API for obtaining the service list**. The **service_id** field indicates the service ID. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A

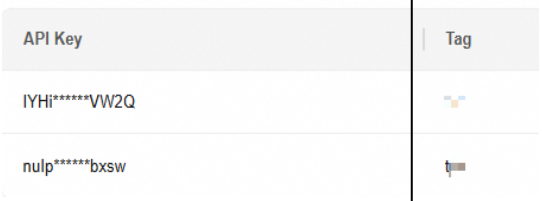
Request Parameters

Table 5-13 Request header parameter

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Table 5-14 Request body parameters

Parameter	Mandatory	Type	Description
service_type	Yes	Integer	**Definition**: Service type. **Constraints**: N/A **Range**: • **1**: User services. Deploy the model service in the **My Services** tab. For details, see **Deploying a Model Service**. • **2**: Built-in services. Deploy the model service in the **Built-in Services** tab. For details, see **Subscribing to a Built-in Service**. **Default Value**: N/A

Parameter	Mandatory	Type	Description
api_keys	No	Array of strings	**Definition**: API key tag list, which is used for filtering. MaaS services support **API key calls**. Go to the **API key management** page to get the API key tag. The **Tag** field in the API key list shows the API key tag.  **Constraints**: N/A **Range**: • Example for obtaining the online experience call data: **api_keys: [""]**. • When obtaining the call data of some API keys, transfer the tags of the corresponding API keys. Example: **api_keys: ["test01","test02"]**. • Example for obtaining the call data of online experience and some API keys: **api_keys: ["test01", "test02", ""]**. • When obtaining all call data, do not use this parameter. **Default Value**: N/A

Parameter	Mandatory	Type	Description
version_id	No	String	**Definition**: Service version ID. If this parameter is not specified, all service versions are queried. To query the service version ID, you can **call the API for querying the service version**. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A
ips	No	Array of strings	**Definition**: IP address list, which indicates the source IP addresses of the clients that have been called. If this parameter is not specified, all IP addresses are queried. To query the IP address, you can **call the API for obtaining the IP address list**. **Constraints**: The value must be in the IP address format. **Range**: N/A **Default Value**: N/A
start_time	Yes	Long	**Definition**: Timestamp of the start time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0** and no greater than the value of **end_time**. **Default Value**: N/A
end_time	Yes	Long	**Definition**: Timestamp of the end time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0**. **Default Value**: N/A

Parameter	Mandatory	Type	Description
time_granularity	Yes	Integer	**Definition**: Time granularity. **Constraints**: The time range (interval between the start time and end time) and time precision must meet the following rules: • For time ranges 0–2 days, precision to minute or hour is supported. • For time ranges 3–7 days, precision to hour or day is supported. • For time ranges 8–30 days, precision to day is supported. **Range**: The value must be an integer ranging from **1** to **3**. • **1**: minute granularity • **2**: hour granularity • **3**: day granularity **Default Value**: N/A
timezone	No	String	**Definition**: Time zone. **Constraints**: N/A **Range**: The value must comply with the IANA time zone specifications, such as Asia/Shanghai and UTC. **Default Value**: **Asia/Shanghai (GMT+8)**
infer_type	Yes	String	**Definition**: Service inference type. **Range**: • **real_time**: real-time inference • **batch**: batch inference (Batch inference is under restricted use. To use it, submit a service ticket.) **Constraints**: N/A **Default Value**: N/A

Parameter	Mandatory	Type	Description
metric	No	String	**Definition**: Metric name. **Constraints**: N/A **Range**: • **rpm**: Query the RPM metric. • **total_token**: Query the total number of tokens. • **prompt_token**: Query the number of input tokens. • **completion_token**: Query the number of output tokens. • **qps**: Query the QPS metric. • In this parameter is not specified, all metrics are queried. **Default Value**: N/A

Response Parameters

Status code: 200

Table 5-15 Response body parameters

Parameter	Type	Description
total	Integer	**Definition**: Total number of returned items. **Range**: N/A
count	Integer	**Definition**: Total number of returned items. **Range**: N/A
items	Array of **DetailStatistics** objects	**Definition**: Detailed metric statistics, which are displayed by time segment. **Range**: N/A

Table 5-16 DetailStatistics

Parameter	Type	Description
time	Long	**Definition**: Timestamp, in milliseconds. **Range**: N/A
request_count	Integer	**Definition**: Number of calls. **Range**: N/A
succ_count	Integer	**Definition**: Number of successful calls. If batch inference is used, this parameter indicates the number of successful inference times. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
error_count	Integer	**Definition**: Number of successful calls. If batch inference is used, this parameter indicates the number of successful inference times. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
error_rate	Double	**Definition**: Call failure rate. If batch inference is used, this parameter indicates the inference failure rate. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: **0** to **1**
total_token	Double	**Definition**: Total number of called tokens (in thousands) **Range**: N/A
avg_total_token	Double	**Definition**: Average number of tokens (in thousands) **Range**: N/A
max_total_token	Double	**Definition**: Maximum number of tokens (in thousands) **Range**: N/A
p50_total_token	Double	**Definition**: Median number of tokens (in thousands) **Range**: N/A

Parameter	Type	Description
p80_total_token	Double	**Definition**: 80th percentile number of tokens (in thousands) **Range**: N/A
p90_total_token	Double	**Definition**: 90th percentile number of tokens (in thousands) **Range**: N/A
p99_total_token	Double	**Definition**: 99th percentile number of tokens (in thousands) **Range**: N/A
prompt_token	Double	**Definition**: Total number of input tokens (in thousands) **Range**: N/A
avg_prompt_token	Double	**Definition**: Average number of input tokens (in thousands) **Range**: N/A
max_prompt_token	Double	**Definition**: Maximum number of input tokens (in thousands) **Range**: N/A
p50_prompt_token	Double	**Definition**: Median number of input tokens (in thousands) **Range**: N/A
p80_prompt_token	Double	**Definition**: 80th percentile number of input tokens (in thousands) **Range**: N/A
p90_prompt_token	Double	**Definition**: 90th percentile number of input tokens (in thousands) **Range**: N/A
p99_prompt_token	Double	**Definition**: 99th percentile number of input tokens (in thousands) **Range**: N/A
completion_token	Double	**Definition**: Total number of output tokens (in thousands) **Range**: N/A
avg_completion_token	Double	**Definition**: Average number of output tokens (in thousands) **Range**: N/A

Parameter	Type	Description
max_completion_token	Double	**Definition**: Maximum number of output tokens (in thousands) **Range**: N/A
p50_completion_token	Double	**Definition**: Median number of output tokens (in thousands) **Range**: N/A
p80_completion_token	Double	**Definition**: 80th percentile number of output tokens (in thousands) **Range**: N/A
p90_completion_token	Double	**Definition**: 90th percentile number of output tokens (in thousands) **Range**: N/A
p99_completion_token	Double	**Definition**: 99th percentile number of output tokens (in thousands) **Range**: N/A
avg_latency	Double	**Definition**: Average E2E latency (ms) **Range**: N/A
max_latency	Double	**Definition**: Maximum E2E latency (ms) **Range**: N/A
p50_latency	Double	**Definition**: Median E2E latency (ms) **Range**: N/A
p80_latency	Double	**Definition**: 80th percentile E2E latency (ms) **Range**: N/A
p90_latency	Double	**Definition**: 90th percentile E2E latency (ms) **Range**: N/A
p99_latency	Double	**Definition**: 99th percentile E2E latency (ms) **Range**: N/A
avg_ttft	Double	**Definition**: Average TTFT (ms). Only streaming responses are counted. **Range**: N/A
max_ttft	Double	**Definition**: Maximum TTFT (ms). Only streaming responses are counted. **Range**: N/A

Parameter	Type	Description
p50_ttft	Double	**Definition**: Median TTFT (ms). Only streaming responses are counted. **Range**: N/A
p80_ttft	Double	**Definition**: 80th percentile TTFT (ms). Only streaming responses are counted. **Range**: N/A
p90_ttft	Double	**Definition**: 90th percentile TTFT (ms). Only streaming responses are counted. **Range**: N/A
p99_ttft	Double	**Definition**: 99th percentile TTFT (ms). Only streaming responses are counted. **Range**: N/A
avg_tpot	Double	**Definition**: Average TPOT (ms). Only streaming responses are counted. **Range**: N/A
max_tpot	Double	**Definition**: Maximum TPOT (ms). Only streaming responses are counted. **Range**: N/A
p50_tpot	Double	**Definition**: Median TPOT (ms). Only streaming responses are counted. **Range**: N/A
p80_tpot	Double	**Definition**: 80th percentile TPOT (ms). Only streaming responses are counted. **Range**: N/A
p90_tpot	Double	**Definition**: 90th percentile TPOT (ms). Only streaming responses are counted. **Range**: N/A
p99_tpot	Double	**Definition**: 99th percentile TPOT (ms). Only streaming responses are counted. **Range**: N/A
rpm	Double	**Definition**: Number of requests processed per minute. **Range**: N/A
tpm	Double	**Definition**: Number of tokens processed per minute (thousand tokens/minute). **Range**: N/A

Parameter	Type	Description
avg_generation_time	Double	**Definition**: Average generation duration (s), that is, the average time generating an image or video. Only video and image generation models are supported. **Range**: N/A
cache_token	Double	**Definition**: Number of cache hits, that is, the total number of tokens that hit the cache in the request. **Range**: N/A
cache_hit_ratio	Double	**Definition**: Cache hit rate, that is, the ratio of cache hit tokens in the request to the total input tokens. **Range**: **0** to **1**
total_token_list	Array of **GradientIndicatorResult** objects	**Definition**: Total token statistics, which includes detailed metric data in the specified period. If the request parameter **metric** is set to **total_token**, statistics about all tokens are returned. **Range**: N/A
prompt_token_list	Array of **GradientIndicatorResult** objects	**Definition**: Input token statistics, which includes detailed metric data in the specified period. If the request parameter **metric** is set to **prompt_token**, statistics about all input tokens are returned. **Range**: N/A
completion_token_list	Array of **GradientIndicatorResult** objects	**Definition**: Output token statistics, which includes detailed metric data in the specified period. If the request parameter **metric** is set to **completion_token**, statistics about all output tokens are returned. **Range**: N/A
rpm_list	Array of **GradientIndicatorResult** objects	**Definition**: RPM details, which includes detailed metric data in the specified period. If the request parameter **metric** is set to **rpm**, RPM details are returned. **Range**: N/A

Parameter	Type	Description
infer_times	Integer	**Definition**: Total number of inference times. This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
completion_tasks_count	Integer	**Definition**: Number of completed tasks. This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
avg_consume_time	Double	**Definition**: Average task handling duration, in minutes. This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
qps	Integer	**Definition**: Number of queries per second (QPS). The peak QPS in the minute is displayed. **Range**: N/A
video_generate_duration	Double	**Definition**: Total duration of the generated video. **Range**: N/A
image_generate_nums	Integer	**Definition**: Total number of generated images. **Range**: N/A

Table 5-17 GradientIndicatorResult

Parameter	Type	Description
name	String	**Definition**: Metric name. **Range**: • RPM • Total number of tokens. • Input token size • Output token size

Parameter	Type	Description
value	Object	**Definition**: Metric value. The int and double types are supported. **Range**: N/A

Status code: 400

Table 5-18 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Query the metric data generated by real-time inference text of the preset service in the last 14 days. The service ID is **4f6d50ec-0e80-4ea0-983b-d0ad1ede7596** and the version ID is **ac73463d-4453-4d62-a3d9-31b627a116b1**.

```
/v1/{{project_id}}/maas/monitoring/4f6d50ec-0e80-4ea0-983b-d0ad1ede7596/show-detail-chart

{
  "service_type" : 2,
  "start_time" : 1768320000000,
  "end_time" : 1769518975857,
  "timezone" : "Asia/Shanghai",
  "time_granularity" : 3,
  "version_id" : "ac73463d-4453-4d62-a3d9-31b627a116b1",
  "infer_type" : "real_time"
}
```

Response Example

Status code: 200

Success response

```
{
  "total" : 14,
  "count" : 14,
  "items" : [ {
    "time" : 1768320000000,
    "request_count" : 35,
    "succ_count" : 13,
    "error_count" : 22,
    "error_rate" : 0.6286,
    "total_token" : 13.149,
```



```
  "avg_total_token" : 1.011,
  "max_total_token" : 3.043,
  "p50_total_token" : 0,
  "p80_total_token" : 0.163,
  "p90_total_token" : 1.647,
  "p99_total_token" : 3.043,
  "prompt_token" : 5.445,
  "avg_prompt_token" : 0.419,
  "max_prompt_token" : 2.747,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0.02,
  "p90_prompt_token" : 0.03,
  "p99_prompt_token" : 2.747,
  "completion_token" : 7.704,
  "avg_completion_token" : 0.593,
  "max_completion_token" : 1.828,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0.133,
  "p90_completion_token" : 1.583,
  "p99_completion_token" : 1.828,
  "avg_latency" : 22811.23,
  "max_latency" : 70615,
  "p50_latency" : 0,
  "p80_latency" : 5839,
  "p90_latency" : 59330,
  "p99_latency" : 70615,
  "avg_ttft" : 522.79,
  "max_ttft" : 1240.61,
  "p50_ttft" : 373.97,
  "p80_ttft" : 634.03,
  "p90_ttft" : 1012.55,
  "p99_ttft" : 1240.61,
  "avg_tpot" : 36.12,
  "max_tpot" : 43,
  "p50_tpot" : 37.27,
  "p80_tpot" : 38.3,
  "p90_tpot" : 39.54,
  "p99_tpot" : 43,
  "rpm" : 0.02,
  "tpm" : 0.009,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,
  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,
  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1768406400000,
  "request_count" : 3,
  "succ_count" : 1,
  "error_count" : 2,
  "error_rate" : 0.6667,
  "total_token" : 1.533,
  "avg_total_token" : 1.533,
  "max_total_token" : 1.533,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 1.533,
  "p99_total_token" : 1.533,
  "prompt_token" : 0.013,
  "avg_prompt_token" : 0.013,
  "max_prompt_token" : 0.013,
```



```
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0.013,
  "p99_prompt_token" : 0.013,
  "completion_token" : 1.52,
  "avg_completion_token" : 1.52,
  "max_completion_token" : 1.52,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 1.52,
  "p99_completion_token" : 1.52,
  "avg_latency" : 56872,
  "max_latency" : 56872,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 56872,
  "p99_latency" : 56872,
  "avg_ttft" : 258.86,
  "max_ttft" : 258.86,
  "p50_ttft" : 258.86,
  "p80_ttft" : 258.86,
  "p90_ttft" : 258.86,
  "p99_ttft" : 258.86,
  "avg_tpot" : 37.27,
  "max_tpot" : 37.27,
  "p50_tpot" : 37.27,
  "p80_tpot" : 37.27,
  "p90_tpot" : 37.27,
  "p99_tpot" : 37.27,
  "rpm" : 0,
  "tpm" : 0.001,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,
  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,
  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1768492800000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
```

```
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,
  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,
  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1768579200000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
```



```
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,
  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,
  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1768665600000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
```



```
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,
  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,
  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1768752000000,
  "request_count" : 3,
  "succ_count" : 3,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0.533,
  "avg_total_token" : 0.178,
  "max_total_token" : 0.228,
  "p50_total_token" : 0.199,
  "p80_total_token" : 0.199,
  "p90_total_token" : 0.228,
  "p99_total_token" : 0.228,
  "prompt_token" : 0.264,
  "avg_prompt_token" : 0.088,
  "max_prompt_token" : 0.139,
  "p50_prompt_token" : 0.105,
  "p80_prompt_token" : 0.105,
  "p90_prompt_token" : 0.139,
  "p99_prompt_token" : 0.139,
  "completion_token" : 0.269,
  "avg_completion_token" : 0.09,
  "max_completion_token" : 0.123,
  "p50_completion_token" : 0.086,
  "p80_completion_token" : 0.086,
  "p90_completion_token" : 0.123,
  "p99_completion_token" : 0.123,
  "avg_latency" : 2962.33,
  "max_latency" : 5112,
  "p50_latency" : 2129,
  "p80_latency" : 2129,
  "p90_latency" : 5112,
  "p99_latency" : 5112,
  "avg_ttft" : 349,
  "max_ttft" : 424.79,
  "p50_ttft" : 422.49,
  "p80_ttft" : 422.49,
  "p90_ttft" : 424.79,
  "p99_ttft" : 424.79,
  "avg_tpot" : 27.02,
  "max_tpot" : 40.27,
  "p50_tpot" : 20.7,
  "p80_tpot" : 20.7,
  "p90_tpot" : 40.27,
  "p99_tpot" : 40.27,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
```



```
  "completion_token_list" : null,
  "rpm_list" : null,

  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,

  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1768838400000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,

  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,
  "qps" : 0,
```



```
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1768924800000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,

  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,

  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1769011200000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
```



```
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,

  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,

  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1769097600000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
```



```
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,

  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,

  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1769184000000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
```



```
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,

  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,

  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1769270400000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
```



```
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,

  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,

  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1769356800000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
```



```
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
  "p50_tpot" : 0,
  "p80_tpot" : 0,
  "p90_tpot" : 0,
  "p99_tpot" : 0,
  "rpm" : 0,
  "tpm" : 0,
  "avg_generation_time" : 0,
  "cache_token" : 0,
  "cache_hit_ratio" : 0,
  "total_token_list" : null,
  "prompt_token_list" : null,
  "completion_token_list" : null,
  "rpm_list" : null,

  "infer_times" : 0,
  "completion_tasks_count" : 0,
  "avg_consume_time" : 0,
  "qps" : 0,
  "video_generate_duration": 0,
  "image_generate_nums": 0
}, {
  "time" : 1769443200000,
  "request_count" : 0,
  "succ_count" : 0,
  "error_count" : 0,
  "error_rate" : 0,
  "total_token" : 0,
  "avg_total_token" : 0,
  "max_total_token" : 0,
  "p50_total_token" : 0,
  "p80_total_token" : 0,
  "p90_total_token" : 0,
  "p99_total_token" : 0,
  "prompt_token" : 0,
  "avg_prompt_token" : 0,
  "max_prompt_token" : 0,
  "p50_prompt_token" : 0,
  "p80_prompt_token" : 0,
  "p90_prompt_token" : 0,
  "p99_prompt_token" : 0,
  "completion_token" : 0,
  "avg_completion_token" : 0,
  "max_completion_token" : 0,
  "p50_completion_token" : 0,
  "p80_completion_token" : 0,
  "p90_completion_token" : 0,
  "p99_completion_token" : 0,
  "avg_latency" : 0,
  "max_latency" : 0,
  "p50_latency" : 0,
  "p80_latency" : 0,
  "p90_latency" : 0,
  "p99_latency" : 0,
  "avg_ttft" : 0,
  "max_ttft" : 0,
  "p50_ttft" : 0,
  "p80_ttft" : 0,
  "p90_ttft" : 0,
  "p99_ttft" : 0,
  "avg_tpot" : 0,
  "max_tpot" : 0,
```



```
    "p50_tpot" : 0,
    "p80_tpot" : 0,
    "p90_tpot" : 0,
    "p99_tpot" : 0,
    "rpm" : 0,
    "tpm" : 0,
    "avg_generation_time" : 0,
    "cache_token" : 0,
    "cache_hit_ratio" : 0,
    "total_token_list" : null,
    "prompt_token_list" : null,
    "completion_token_list" : null,
    "rpm_list" : null,

    "infer_times" : 0,
    "completion_tasks_count" : 0,
    "avg_consume_time" : 0,
    "qps" : 0,
    "video_generate_duration": 0,
    "image_generate_nums": 0
  } ]
}
```


Status code: 400

Failure response


```
{
  "error_code" : "ModelArts.0104",
  "error_msg" : "Inference type realtime is invalid. The inference type must be real_time or batch."
}
```


Status Codes

Status Code	Description
200	Success response
400	Failure response

Error Codes

For details, see **Error Codes**.

5.4 Obtaining Service Error Details

Function

This API is used to retrieve detailed error data for services to review failure information, such as error codes, occurrences, and error messages. Data is retained for 30 days only.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 1,000 within one minute.

- User request limit: The number of requests for this API from a single user cannot exceed 200 within one minute.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 60 seconds and try again.

URI

POST /v1/{project_id}/maas/monitoring/{service_id}/list-errors

Table 5-19 URI parameters

Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A
service_id	Yes	String	**Definition**: Service IDs to be queried. Services are filtered based on the input service ID list. If this parameter is not specified, all service names corresponding to the IDs are returned. You can obtain the service ID from the response body during service creation, or **call the API for obtaining the service list**. The **service_id** field indicates the service ID. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A

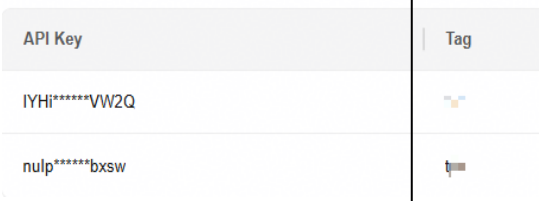
Request Parameters

Table 5-20 Request header parameters

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A
Accept-Language	No	String	**Definition**: Specify the expected language of the response. The server returns information or data in the specified language. **Constraints**: The value is in the format of *<language-code>-<region-code>*, for example, **zh** (Chinese), **en-US** (English-US). Multiple language priorities are supported. Use commas (,) to separate them. You can use the **q** parameter (**0** to **1**, default value **1**) to specify the weight. Example: **Accept-Language: zh-CN,en-US;q=0.9**. **Range**: N/A (The client only needs to transfer the standard language label, such as **zh-CN** and **en-US**.) **Default Value**: N/A

Table 5-21 Request body parameters

Parameter	Mandatory	Type	Description
service_type	Yes	Integer	**Definition**: Service type. **Constraints**: N/A **Range**: • **1**: User services. Deploy the model service in the **My Services** tab. For details, see **Deploying a Model Service**. • **2**: Built-in services. Deploy the model service in the **Built-in Services** tab. For details, see **Subscribing to a Built-in Service**. **Default Value**: N/A

Parameter	Mandatory	Type	Description
api_keys	No	Array of strings	**Definition**: API key tag list, which is used for filtering. MaaS services support **API key calls**. Go to the **API key management** page to get the API key tag. The **Tag** field in the API key list shows the API key tag.  **Constraints**: N/A **Range**: • Example for obtaining the online experience call data: **api_keys: [""]**. • When obtaining the call data of some API keys, transfer the tags of the corresponding API keys. Example: **api_keys: ["test01","test02"]**. • Example for obtaining the call data of online experience and some API keys: **api_keys: ["test01", "test02", ""]**. • When obtaining all call data, do not use this parameter. **Default Value**: N/A

Parameter	Mandatory	Type	Description
version_id	No	String	**Definition**: Service version ID. If this parameter is not specified, all service versions are queried. To query the service version ID, you can **call the API for querying the service version**. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A
ips	No	Array of strings	**Definition**: IP address list, which indicates the source IP addresses of the clients that have been called. If this parameter is not specified, all IP addresses are queried. To query the IP address, you can **call the API for obtaining the IP address list**. **Constraints**: The value must be in the IP address format. **Range**: N/A **Default Value**: N/A
start_time	Yes	Long	**Definition**: Timestamp of the start time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0** and no greater than the value of **end_time**. **Default Value**: N/A
end_time	Yes	Long	**Definition**: Timestamp of the end time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0**. **Default Value**: N/A

Parameter	Mandatory	Type	Description
timezone	No	String	**Definition**: Time zone. **Constraints**: N/A **Range**: The value must comply with the IANA time zone specifications, such as Asia/Shanghai and UTC. **Default Value**: **Asia/Shanghai (GMT+8)**
infer_type	Yes	String	**Definition**: Service inference type. **Range**: • **real_time**: real-time inference • **batch**: batch inference (Batch inference is under restricted use. To use it, submit a service ticket.) **Constraints**: N/A **Default Value**: N/A

Response Parameters

Status code: 200

Table 5-22 Response body parameters

Parameter	Type	Description
total	Integer	**Definition**: Error code type, which can be 4*xx* or 5*xx*. **Range**: The value is fixed to **2**.
count	Integer	**Definition**: Error code type, which can be 4*xx* or 5*xx*. **Range**: The value is fixed to **2**.
items	Array of **ErrorsStatisticsItem** objects	**Definition**: List of error code details. **Range**: N/A

Table 5-23 ErrorsStatisticsItem

Parameter	Type	Description
error_code	String	**Definition**: Error code. **Range**: **4***xx* or **5***xx*.
error_count	Integer	**Definition**: Number of errors. **Range**: N/A
ratio	Double	**Definition**: Number of errors of this code to the total number of errors. **Range**: **0** to **1**
error_desc	String	**Definition**: Error description. **Range**: N/A
details	Array of **ErrorsStatisticsItem** objects	**Definition**: Information about a specific 4*xx*/5*xx* error, including the error code, number of errors, error percentage, and error description. **Range**: The error codes and messages are as follows: • **401**: Insufficient permission. Check your authentication information. • **403**: Non-compliant content is requested or generated. • **404**: Resource not found. The request path may be incorrect. • **413**: The request body is too large and is rejected by the server. • **429**: Request traffic is limited. • **499**: The client proactively disables the connection or cancels the request. • **500**: An unknown error occurred on the server. • **503**: No inference service is available at the bottom layer. • **504**: The request time exceeds the maximum limit.

Status code: 400

Table 5-24 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Query the error details of real-time inference for a user service within 14 days. The service ID is **69fefc89-7d71-4936-bf4d-c0c33616558d**, and the version ID is **943cb312-2dac-4402-b9a9-d99de4861cb2**.

```
/v1/{{project_id}}/maas/monitoring/69fefc89-7d71-4936-bf4d-c0c33616558d/list-errors

{
  "service_type" : 2,
  "start_time" : 1768320000000,
  "end_time" : 1769523118705,
  "timezone" : "Asia/Shanghai",
  "version_id" : "943cb312-2dac-4402-b9a9-d99de4861cb2",
  "infer_type" : "real_time"
}
```

Response Example

Status code: 200

Success response

```
{
  "total" : 2,
  "count" : 2,
  "items" : [ {
    "error_code" : "4xx",
    "error_count" : 0,
    "ratio" : 0,
    "error_desc" : "Client Errors: Invalid request (for example, incorrect format or insufficient permission). This error usually occurs on the client.",
    "details" : [ ]
  }, {
    "error_code" : "5xx",
    "error_count" : 20,
    "ratio" : 1,
    "error_desc" : "Server Errors: Internal error during request processing on the server.",
    "details" : [ {
      "error_code" : "500",
      "error_count" : 20,
      "ratio" : 1,
      "error_desc" : "An unknown error occurred on the server.",
      "details" : [ ]
    } ]
  } ]
}
```


Status code: 400

Failure response

```
{
  "error_code" : "ModelArts.0104",
  "error_msg" : "Invalid parameter. Issue: The end time cannot be earlier than the start time."
}
```

Status Codes

Status Code	Description
200	Success response
400	Failure response

Error Codes

For details, see **Error Codes**.

5.5 Obtaining the Service List

Function

This API is used to obtain the service name based on the service ID.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 1,000 within one minute.
- User request limit: The number of requests for this API from a single user cannot exceed 200 within one minute.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 60 seconds and try again.

URI

POST /v1/{project_id}/maas/monitoring/list-services

Table 5-25 URI parameter

Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A

Request Parameters

Table 5-26 Request header parameter

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Table 5-27 Request body parameters

Parameter	Mandatory	Type	Description
service_type	Yes	Integer	**Definition**: Service type. **Constraints**: N/A **Range**: • **1**: User services. Deploy the model service in the **My Services** tab. For details, see **Deploying a Model Service**. • **2**: Built-in services. Deploy the model service in the **Built-in Services** tab. For details, see **Subscribing to a Built-in Service**. **Default Value**: N/A
service_ids	No	Array of strings	**Definition**: Service IDs to be queried. Services are filtered based on the input service ID list. If this parameter is not specified, all service names corresponding to the IDs are returned. You can obtain the service ID from the response body during service creation, or **call the API for obtaining the service list**. The **service_id** field indicates the service ID. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A
limit	No	Integer	**Definition**: Number of records on each page, that is, maximum number of returned records. **Constraints**: N/A **Range**: **0** to **100** **Default Value**: **10**.

Parameter	Mandatory	Type	Description
offset	No	Integer	**Definition**: Pagination offset, which indicates the number of records to be skipped. **Constraints**: N/A **Range**: The value must be greater than or equal to 0. **Default Value**: **0**

Response Parameters

Status code: 200

Table 5-28 Response body parameters

Parameter	Type	Description
total	Integer	**Definition**: Total number of queried services. **Range**: N/A
count	Integer	**Definition**: Number of data records returned on the current page. **Range**: N/A
items	Array of **ListServiceItem** objects	**Definition**: Service list. **Range**: N/A

Table 5-29 ListServiceItem

Parameter	Type	Description
service_id	String	**Definition**: Service ID. **Range**: N/A
service_name	String	**Definition**: Service name. **Range**: N/A

Status code: 400

Table 5-30 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Query the name of the preset service whose ID is **08d1b2ad-da2e-4a69-b71a-eedb9b120121**.

```
/v1/{project_id}/maas/monitoring/list-services

{
  "service_ids" : [ "08d1b2ad-da2e-4a69-b71a-eedb9b120121" ],
  "service_type" : 2
}
```

Response Example

Status code: 200

Success response

```
{
  "total" : 1,
  "count" : 1,
  "items" : [ {
    "service_id" : "4f6d50ec-0e80-4ea0-983b-d0ad1ede7596",
    "service_name" : "deepseek.v3.1"
  } ]
}
```

Status code: 400

Failure response

```
{
  "error_code" : "common.00000400",
  "error_msg" : "The value of field limit must range from 0 to 100."
}
```

Status Codes

Status Code	Description
200	Success response
400	Failure response

Error Codes

For details, see **Error Codes**.

5.6 Querying the IP Address List

Function

This API is used to obtain the IP addresses.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 1,000 within one minute.
- User request limit: The number of requests for this API from a single user cannot exceed 200 within one minute.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 60 seconds and try again.

URI

POST /v1/{project_id}/maas/monitoring/source-ips

Table 5-31 URI parameter

Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A

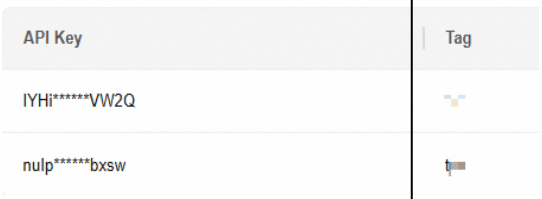
Request Parameters

Table 5-32 Request header parameter

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Table 5-33 Request body parameters

Parameter	Mandatory	Type	Description
service_type	Yes	Integer	**Definition**: Service type. **Constraints**: N/A **Range**: • **1**: User services. Deploy the model service in the **My Services** tab. For details, see **Deploying a Model Service**. • **2**: Built-in services. Deploy the model service in the **Built-in Services** tab. For details, see **Subscribing to a Built-in Service**. **Default Value**: N/A
limit	Yes	Integer	**Definition**: Number of records on each page, that is, maximum number of returned records. **Constraints**: N/A **Range**: Greater than 0. **Default Value**: N/A

Parameter	Mandatory	Type	Description
offset	No	Integer	**Definition**: Pagination offset, which indicates the number of records to be skipped. **Constraints**: N/A **Range**: The value must be greater than or equal to 0. **Default Value**: N/A
api_keys	No	Array of strings	**Definition**: API key tag list, which is used for filtering. MaaS services support **API key calls**. Go to the **API key management** page to get the API key tag. The **Tag** field in the API key list shows the API key tag.  **Constraints**: N/A **Range**: • Example for obtaining the online experience call data: **api_keys: [""]**. • When obtaining the call data of some API keys, transfer the tags of the corresponding API keys. Example: **api_keys: ["test01","test02"]**. • Example for obtaining the call data of online experience and some API keys: **api_keys: ["test01", "test02", ""]**. • When obtaining all call data, do not use this parameter. **Default Value**: N/A

Parameter	Mandatory	Type	Description
ip_search	No	String	**Definition**: Fuzzy query of IP addresses. If this parameter is specified, IP addresses starting with the parameter value are queried. **Constraints**: N/A **Range**: IPv4 address prefix, for example, **192.168** and **10.0**. **Default Value**: N/A
start_time	Yes	Long	**Definition**: Timestamp of the start time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0** and no greater than the value of **end_time**. **Default Value**: N/A
end_time	Yes	Long	**Definition**: Timestamp of the end time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0**. **Default Value**: N/A
service_id	No	String	**Definition**: Service ID. You can obtain the service ID from the response body during service creation, or **call the API for obtaining the service list**. The **service_id** field indicates the service ID. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A

Parameter	Mandatory	Type	Description
timezone	No	String	**Definition**: Time zone. **Constraints**: N/A **Range**: The value must comply with the IANA time zone specifications, such as Asia/Shanghai and UTC. **Default Value**: **Asia/Shanghai (GMT+8)**

Response Parameters

Status code: 200

Table 5-34 Response body parameters

Parameter	Type	Description
total	Integer	**Definition**: Total number of queried data records. **Range**: N/A
items	Array of strings	**Definition**: IP address list. **Range**: N/A
count	Integer	**Definition**: Maximum number of data records returned on each page. **Range**: The value is that of the input parameter **limit**.

Status code: 400

Table 5-35 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Perform fuzzy query on preset services whose IP addresses start with 192 in the GMT+8 time zone. Ten records are displayed on each page, starting from the first record.

```
/v1/{project_id}/maas/monitoring/source-ips

{
  "limit" : 10,
  "offset" : 0,
  "start_time" : 1768406400000,
  "end_time" : 1769529757467,
  "timezone" : "Asia/Shanghai",
  "service_type" : 2,
  "ip_search" : "192"
}
```

Response Example

Status code: 200

Success response

```
{
  "total" : 2,
  "items" : [ "192.168.1.148", "192.168.4.99" ],
  "count" : 10
}
```

Status code: 400

Failure response

```
{
  "error_code" : "ModelArts.0104",
  "error_msg" : "Invalid parameter. Issue: The end time cannot be earlier than the start time."
}
```

Status Codes

Status Code	Description
200	Success response
400	Failure response

Error Codes

For details, see **Error Codes**.

5.7 Querying Resource Monitoring Metric Details

Function

This API is used to obtain the resource metric information of **My Services** in the MaaS real-time inference module. Data is retained for 30 days only.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 1,000 within one minute.
- User request limit: The number of requests for this API from a single user cannot exceed 200 within one minute.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 60 seconds and try again.

URI

GET /v1/{project_id}/maas/monitoring/{service_id}/detail-statistics

Table 5-36 URI parameters

Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A
service_id	Yes	String	**Definition**: Service IDs to be queried. Services are filtered based on the input service ID list. If this parameter is not specified, all service names corresponding to the IDs are returned. You can obtain the service ID from the response body during service creation, or **call the API for obtaining the service list**. The **service_id** field indicates the service ID. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A

Table 5-37 Query parameters

Parameter	Mandatory	Type	Description
start_time	Yes	Long	**Definition**: Timestamp of the start time, in milliseconds. **Constraints**: N/A **Range**: The value must be greater than **0** and less than or equal to the value of **end_time**. The end time must be at least 30 days from the start time. **Default Value**: N/A
end_time	Yes	Long	**Definition**: Timestamp of the end time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0**. **Default Value**: N/A

Parameter	Mandatory	Type	Description
metric_name	Yes	String	**Definition**: Metric name. **Constraints**: N/A **Range**: • **npu_util**: NPU compute usage • **cpu_usage**: CPU usage • **mem_usage**: memory usage • **npu_mem_usage**: NPU memory usage • **disk_read_rate**: disk read rate • **disk_write_rate**: disk write rate • **recv_bytes_rate**: downlink rate • **send_bytes_rate**: uplink rate • **running_task**: number of running requests • **pending_task**: number of pending requests • **kv_cache_usage**: KV cache usage • **prompt_tps**: input TPS • **completion_tps**: output TPS **Default Value**: N/A
timezone	No	String	**Definition**: Time zone. **Constraints**: N/A **Range**: The value must comply with the IANA time zone specifications, such as Asia/Shanghai and UTC. **Default Value**: **Asia/Shanghai (GMT+8)**

Parameter	Mandatory	Type	Description
model_id	Yes	String	**Definition**: Model ID. To obtain the model ID, log in to the MaaS console. In the navigation pane on the left, choose **Model Inference** > **Real-Time Inference**. In the **My Services** tab, click the service name. On the displayed details page, press **F12** and access the **Network** tab. Then, view the value of **model_id** in the response body of the models API. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Request Parameters

Table 5-38 Request header parameter

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Response Parameters

Status code: 200

Table 5-39 Response body parameters

Parameter	Type	Description
points	Array of **DataPoint** objects	**Definition**: Data point. **Range**: N/A

Parameter	Type	Description
metric_name	String	**Definition**: Metric name. **Range**: • **npu_util**: NPU compute usage • **cpu_usage**: CPU usage • **mem_usage**: memory usage • **npu_mem_usage**: NPU memory usage • **disk_read_rate**: disk read rate • **disk_write_rate**: disk write rate • **recv_bytes_rate**: downlink rate • **send_bytes_rate**: uplink rate • **running_task**: number of running requests • **pending_task**: number of pending requests • **kv_cache_usage**: KV cache usage • **prompt_tps**: input TPS • **completion_tps**: output TPS
unit	String	**Definition**: Metric unit. **Range**: N/A

Table 5-40 DataPoint

Parameter	Type	Description
val	Double	**Definition**: Metric value. **Range**: N/A
time	Long	**Definition**: Timestamp, in milliseconds. **Range**: N/A

Status code: 400

Table 5-41 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Query the output TPS of the service whose service ID is *{<service-ID>}* and model ID is **29f474eb-67bc-4a7c-9771-c98c568c1c8c** from 1769532157287 to 1769535757287.

```
/v1/{project_id}/maas/monitoring/{service_id}/detail-statistics?
start_time=1769532157287&end_time=1769535757287&metric_name=completion_tps&model_id=29f474eb-67bc-4a7c-9771-c98c568c1c8c&timezone=Asia/Shanghai
```

Response Example

Status code: 200

Success response

```
{
  "points" : [ ],
  "metric_name" : "running_task",
  "unit" : "num"
}
```

Status code: 400

Failure response

```
{
  "error_msg" : "Database error.",
  "error_code" : "ModelArts.2010"
}
```

Status Codes

Status Code	Description
200	Success response
400	Failure response

Error Codes

For details, see **Error Codes**.

5.8 Querying the Calling Data of a Service Version

Function

This API is used to query all versions of a service and their corresponding monitoring metric data. Data is retained for 30 days only.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 1,000 within one minute.
- User request limit: The number of requests for this API from a single user cannot exceed 200 within one minute.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 60 seconds and try again.

URI

POST /v1/{project_id}/maas/monitoring/{service_id}/list-version-statistics

Table 5-42 URI parameters

Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A

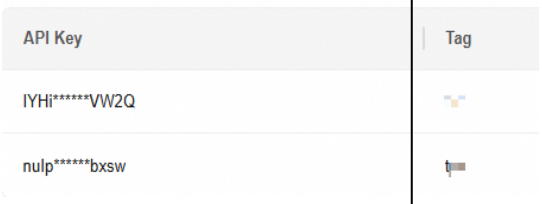
Parameter	Mandatory	Type	Description
service_id	Yes	String	**Definition**: Service IDs to be queried. Services are filtered based on the input service ID list. If this parameter is not specified, all service names corresponding to the IDs are returned. You can obtain the service ID from the response body during service creation, or **call the API for obtaining the service list**. The **service_id** field indicates the service ID. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A

Request Parameters

Table 5-43 Request header parameter

Parameter	Mandatory	Type	Description
X-Auth-Token	No	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Table 5-44 Request body parameters

Parameter	Mandatory	Type	Description
api_keys	No	Array of strings	**Definition**: API key tag list, which is used for filtering. MaaS services support **API key calls**. Go to the **API key management** page to get the API key tag. The **Tag** field in the API key list shows the API key tag.  **Constraints**: N/A **Range**: • Example for obtaining the online experience call data: **api_keys: [""]**. • When obtaining the call data of some API keys, transfer the tags of the corresponding API keys. Example: **api_keys: ["test01","test02"]**. • Example for obtaining the call data of online experience and some API keys: **api_keys: ["test01", "test02", ""]**. • When obtaining all call data, do not use this parameter. **Default Value**: N/A

Parameter	Mandatory	Type	Description
ips	No	Array of strings	**Definition**: IP address list, which indicates the source IP addresses of the clients that have been called. If this parameter is not specified, all IP addresses are queried. To query the IP address, you can **call the API for obtaining the IP address list**. **Constraints**: The value must be in the IP address format. **Range**: N/A **Default Value**: N/A
start_time	Yes	Long	**Definition**: Timestamp of the start time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0** and no greater than the value of **end_time**. **Default Value**: N/A
end_time	Yes	Long	**Definition**: Timestamp of the end time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0**. **Default Value**: N/A
timezone	No	String	**Definition**: Time zone. **Constraints**: N/A **Range**: The value must comply with the IANA time zone specifications, such as Asia/Shanghai and UTC. **Default Value**: **Asia/Shanghai (GMT+8)**

Parameter	Mandatory	Type	Description
infer_type	Yes	String	**Definition**: Service inference type. **Range**: • **real_time**: real-time inference • **batch**: batch inference (Batch inference is under restricted use. To use it, submit a service ticket.) **Constraints**: N/A **Default Value**: N/A

Response Parameters

Status code: 200

Table 5-45 Response body parameters

Parameter	Type	Description
total	Integer	**Definition**: Total number of queried versions. **Range**: N/A
count	Integer	**Definition**: Total number of queried versions. **Range**: N/A
items	Array of **VersionStatItem** objects	**Definition**: Version information list. **Range**: N/A

Table 5-46 VersionStatItem

Parameter	Type	Description
service_id	String	**Definition**: Service ID. **Range**: N/A
version_id	String	**Definition**: Version ID. **Range**: N/A
version_name	String	**Definition**: Version name. **Range**: N/A

Parameter	Type	Description
request_count	Integer	**Definition**: Number of calls. **Range**: N/A
error_count	Integer	**Definition**: Number of failed calls. If batch inference is used, this parameter indicates the number of failed inference times. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
error_rate	Double	**Definition**: Call failure rate. If batch inference is used, this parameter indicates the inference failure rate. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: **0** to **1**
total_token	Double	**Definition**: Total number of called tokens, in thousands. If batch inference is used, this parameter indicates the total number of tokens. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
prompt_token	Double	**Definition**: Number of input tokens **Range**: N/A
completion_token	Double	**Definition**: Number of output tokens **Range**: N/A
avg_latency	Double	**Definition**: Average response time (ms) **Range**: N/A
avg_ttft	Double	**Definition**: Average TTFT (ms) **Range**: N/A
avg_tpot	Double	**Definition**: Average TPOT (ms) **Range**: N/A
completion_tasks_count	Integer	**Definition**: Number of completed tasks **Range**: N/A

Parameter	Type	Description
infer_times	Integer	**Definition**: Total number of inference times. (This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket.) **Range**: N/A
avg_consume_time	Double	**Definition**: Average task handling duration, in minutes. This parameter is related to batch inference. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
avg_generation_time	Double	**Definition**: Average generation duration (s), that is, the average time generating an image or video. Only video and image generation models are supported. **Range**: N/A
cache_token	Double	**Definition**: Number of cache hit tokens (in thousands). **Range**: N/A
cache_hit_ratio	Double	**Definition**: Cache hit rate, which is number of cache hit tokens/number of input tokens. **Range**: **0** to **1**
scc_count	Integer	**Definition**: Number of successful calls. If batch inference is used, this parameter indicates the number of successful inference times. However, batch inference is under restricted use. To use it, submit a service ticket. **Range**: N/A
video_generate_duration	Double	**Definition**: Total duration of the generated video. **Range**: N/A
image_generate_nums	Integer	**Definition**: Total number of generated images. **Range**: N/A

Status code: 400

Table 5-47 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Query the version information and metric data of the real-time inference service whose service ID is *{service-ID}* in GMT+8.

```
/v1/{project_id}/maas/monitoring/{service_id}/list-version-statistics

{
  "start_time" : 1768406400000,
  "end_time" : 1769531307195,
  "timezone" : "Asia/Shanghai",
  "infer_type" : "real_time"
}
```

Response Example

Status code: 200

Success response

```
{
  "total" : 1,
  "count" : 1,
  "items" : [ {
    "service_id" : "4f6d50ec-0e80-4ea0-983b-d0ad1ede7596",
    "version_id" : "ac73463d-4453-4d62-a3d9-31b627a116b1",
    "version_name" : "Qwen2-7B-3.1",
    "request_count" : 6,
    "error_count" : 2,
    "error_rate" : 0.3333,
    "total_token" : 2.066,
    "prompt_token" : 0.277,
    "completion_token" : 1.789,
    "avg_latency" : 16439.75,
    "avg_ttft" : 326.46,
    "avg_tpot" : 29.58,
    "completion_tasks_count" : 0,
    "infer_times" : 0,
    "avg_consume_time" : 0,
    "avg_generation_time" : 0,
    "cache_token" : 0,
    "cache_hit_ratio" : 0,
    "scc_count" : 0,
    "video_generate_duration": 0,
    "image_generate_nums": 0
  } ]
}
```

Status code: 400

Failure response

```
{
  "error_code" : "ModelArts.0104",
  "error_msg" : "Invalid parameter. Issue: The end time cannot be earlier than the start time."
}
```

Status Codes

Status Code	Description
200	Success response
400	Failure response

Error Codes

For details, see **Error Codes**.

5.9 Obtaining the Metrics Supported by Different Model Types

Function

This API is used to obtain the metrics supported by different model types.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 1,000 within one minute.
- User request limit: The number of requests for this API from a single user cannot exceed 200 within one minute.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 60 seconds and try again.

URI

GET /v1/{project_id}/maas/monitoring/generation-supported-metrics

Table 5-48 URI parameter

Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A

Request Parameters

Table 5-49 Request header parameter

Parameter	Mandatory	Type	Description
x-auth-token	Yes	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Response Parameters

Status code: 200

Table 5-50 Response body parameter

Parameter	Type	Description
[Array elements]	Array of **GenerationMetricsItem** objects	Result item corresponding to the model type in GenerationSupportMetricsResponse.

Table 5-51 GenerationMetricsItem

Parameter	Type	Description
type	String	**Definition**: Model type. **Range**: • **Text Generation**: text generation model • **Video Generation**: video generation model • **Image Generation**: image generation model • **Vector Model**: text vectorization • **Embedding**: embedding model • **Image Understanding**: image understanding model • **Rerank**: reranking model
metrics	Array of strings	**Definition**: Monitoring metrics supported of this model type. **Range**: N/A
desc_zh	String	**Definition**: Description in Chinese. **Range**: N/A
desc_en	String	**Definition**: Description in English. **Range**: N/A

Status code: 400

Table 5-52 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Query the calling statistics metrics supported by each type of model.

```
/v1/{project_id}/maas/monitoring/generation-supported-metrics
```


Response Example

Status code: 200

Success response

```
[ {
  "type" : "Text Generation",
  "metrics" : [ "request_count", "succ_2xx_count", "error_count", "req_count4xx5xx", "error_rate",
"total_token", "avg_total_token", "max_total_token", "p50_total_token", "p80_total_token",
"p90_total_token", "p99_total_token", "prompt_token", "completion_token", "avg_prompt_token",
"p50_prompt_token", "p80_prompt_token", "p90_prompt_token", "p99_prompt_token",
"max_prompt_token", "avg_completion_token", "p50_completion_token", "p80_completion_token",
"p90_completion_token", "p99_completion_token", "max_completion_token", "avg_latency", "rpm", "tpm",
"avg_ttft", "p50_ttft", "p80_ttft", "p90_ttft", "p99_ttft", "max_ttft", "avg_tpot", "p50_tpot", "p80_tpot",
"p90_tpot", "p99_tpot", "max_tpot", "cache_token", "cache_hit_ratio" ],
  "desc_zh" : "Text generation model",
  "desc_en" : "Text generation model"
}]
```

Status code: 400

Failure response

```
{
  "error_msg" : "The project ID in the request does not match that in the token.",
  "error_code" : "ModelArts.0210"
}
```

Status Codes

Status Code	Description
200	Success response
400	Failure response

Error Codes

For details, see **Error Codes**.

5.10 Obtaining Time-based Service Error Code Statistics

Function

This API is used to display the chronological distribution of service error code statistics. Data is retained for 30 days only.

Constraints

- Region restrictions: Only the CN-Hong Kong region is supported.
- API rate limiting: The total number of requests for this API from all users cannot exceed 80 within 20 seconds.

- User request limit: The number of requests for this API from a single user cannot exceed 1 within 20 seconds.
- Rate-limiting response: When the rate limit is exceeded, the API returns the HTTP status code 429 "Too Many Requests".
- Retry suggestion: If the rate limit is exceeded, wait 20 seconds and try again.

URI

POST /v1/{project_id}/maas/monitoring/{service_id}/error-code-chart

Table 5-53 URI parameters

Parameter	Mandatory	Type	Description
project_id	Yes	String	**Definition**: Project ID. For details about how to obtain the project ID, see **Obtaining a Project ID and Name**. **Constraints**: N/A **Range**: The value can contain 32 characters. Only lowercase letters and digits are allowed. **Default Value**: N/A
service_id	Yes	String	**Definition**: Service IDs to be queried. Services are filtered based on the input service ID list. If this parameter is not specified, all service names corresponding to the IDs are returned. You can obtain the service ID from the response body during service creation, or **call the API for obtaining the service list**. The **service_id** field indicates the service ID. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A

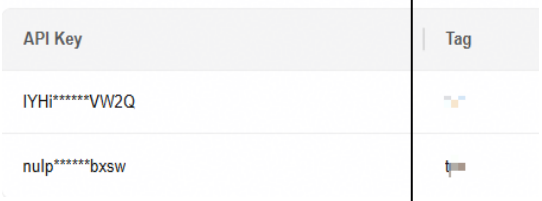
Request Parameters

Table 5-54 Request header parameter

Parameter	Mandatory	Type	Description
X-Auth-Token	Yes	String	**Definition**: User token. It can be obtained by calling the IAM API that is used to obtain a user token. The value of **X-Subject-Token** in the response header is the user token. For details, see **Authentication**. **Constraints**: N/A **Range**: N/A **Default Value**: N/A

Table 5-55 Request body parameters

Parameter	Mandatory	Type	Description
service_type	Yes	Integer	**Definition**: Service type. **Constraints**: N/A **Range**: • **1**: User services. Deploy the model service in the **My Services** tab. For details, see **Deploying a Model Service**. • **2**: Built-in services. Deploy the model service in the **Built-in Services** tab. For details, see **Subscribing to a Built-in Service**. **Default Value**: N/A

Parameter	Mandatory	Type	Description
api_keys	No	Array of strings	**Definition**: API key tag list, which is used for filtering. MaaS services support **API key calls**. Go to the **API key management** page to get the API key tag. The **Tag** field in the API key list shows the API key tag.  **Constraints**: N/A **Range**: • Example for obtaining the online experience call data: **api_keys: [""]**. • When obtaining the call data of some API keys, transfer the tags of the corresponding API keys. Example: **api_keys: ["test01","test02"]**. • Example for obtaining the call data of online experience and some API keys: **api_keys: ["test01", "test02", ""]**. • When obtaining all call data, do not use this parameter. **Default Value**: N/A

Parameter	Mandatory	Type	Description
ips	No	Array of strings	**Definition**: IP address list, which indicates the source IP addresses of the clients that have been called. If this parameter is not specified, all IP addresses are queried. To query the IP address, you can **call the API for obtaining the IP address list**. **Constraints**: The value must be in the IP address format. **Range**: N/A **Default Value**: N/A
start_time	Yes	Long	**Definition**: Timestamp of the start time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0** and no greater than the value of **end_time**. **Default Value**: N/A
end_time	Yes	Long	**Definition**: Timestamp of the end time, in milliseconds. **Constraints**: The end time must be within 30 days from the start time. **Range**: The value must be greater than **0**. **Default Value**: N/A
timezone	No	String	**Definition**: Time zone. **Constraints**: N/A **Range**: The value must comply with the IANA time zone specifications, such as Asia/Shanghai and UTC. **Default Value**: **Asia/Shanghai (GMT+8)**

Parameter	Mandatory	Type	Description
infer_type	Yes	String	**Definition**: Service inference type. **Range**: • **real_time**: real-time inference • **batch**: batch inference (Batch inference is under restricted use. To use it, submit a service ticket.) **Constraints**: N/A **Default Value**: N/A
error_code_type	No	String	**Definition**: Error code type. **Constraints**: N/A **Range**: • **4xx**: Error codes starting with 4. • **5xx**: Error codes starting with 5. If this parameter is left empty or not specified, all error codes starting with 4 and 5 are displayed. **Default Value**: N/A

Parameter	Mandatory	Type	Description
time_granularity	Yes	Integer	**Definition**: Time granularity. **Constraints**: The time range (interval between the start time and end time) and time precision must meet the following rules: • For time ranges 0–2 days, precision to minute or hour is supported. • For time ranges 3–7 days, precision to hour or day is supported. • For time ranges 8–30 days, precision to day is supported. **Range**: The value must be an integer ranging from **1** to **3**. • **1**: minute granularity • **2**: hour granularity • **3**: day granularity **Default Value**: N/A
version_id	No	String	**Definition**: Service version ID, which is used for filtering. If this parameter is not specified, data of all versions is queried. To query the service version ID, you can **call the API for querying the service version**. **Constraints**: N/A **Range**: The value can contain 1 to 128 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed. **Default Value**: N/A

Response Parameters

Status code: 200

Table 5-56 Response body parameters

Parameter	Type	Description
total	Integer	**Definition**: Total number of errors. **Range**: N/A
count	Integer	**Definition**: Total number of errors. **Range**: N/A
list_4xx	Array of **ErrorCodeCount** objects	**Definition**: 4xx error details. **Range**: N/A
list_5xx	Array of **ErrorCodeCount** objects	**Definition**: 5xx error details. **Range**: N/A

Table 5-57 ErrorCodeCount

Parameter	Type	Description
error_code	String	**Definition**: 4xx or 5xx error codes. **Range**: The error codes and messages are as follows: • **401**: Insufficient permission. Check your authentication information. • **403**: Non-compliant content is requested or generated. • **404**: Resource not found. The request path may be incorrect. • **413**: The request body is too large and is rejected by the server. • **429**: Request traffic is limited. • **499**: The client proactively disables the connection or cancels the request. • **500**: An unknown error occurred on the server. • **503**: No inference service is available at the bottom layer. • **504**: Request timed out.
list	Array of **TimestampErrorCnt** objects	**Definition**: Error code occurrences in each time period, including the timestamp in milliseconds and the number of errors. **Range**: N/A

Table 5-58 TimestampErrorCnt

Parameter	Type	Description
time	Long	**Definition**: Timestamp, in milliseconds. **Range**: N/A
count	Long	**Definition**: Number of error codes in the specified period. **Range**: N/A

Status code: 400

Table 5-59 Response body parameters

Parameter	Type	Description
error_code	String	**Definition**: Error code, which identifies the error type. For details, see **MaaS Error Codes**. **Range**: N/A
error_msg	String	**Definition**: Error description. For details, see **MaaS Error Codes**. **Range**: N/A

Request Example

Query the metric data generated by real-time inference text of the preset service in the last 14 days. The service ID is **4f6d50ec-0e80-4ea0-983b-d0ad1ede7596** and the version ID is **ac73463d-4453-4d62-a3d9-31b627a116b1**.

```
/v1/{{project_id}}/maas/monitoring/4f6d50ec-0e80-4ea0-983b-d0ad1ede7596/error-code-chart

{
  "service_type" : 2,
  "start_time" : 1768406400000,
  "end_time" : 1769532163224,
  "timezone" : "Asia/Shanghai",
  "time_granularity" : 3,
  "infer_type" : "real_time"
}
```

Response Example

Status code: 200

Success response

```
{
  "total" : 2,
  "count" : 2,
  "list_4xx" : [ {
```



```
  "error_code" : "400",
  "list" : [ {
    "time" : 1768406400000,
    "count" : 2
  }, {
    "time" : 1768492800000,
    "count" : 0
  }, {
    "time" : 1768579200000,
    "count" : 0
  }, {
    "time" : 1768665600000,
    "count" : 0
  }, {
    "time" : 1768752000000,
    "count" : 0
  }, {
    "time" : 1768838400000,
    "count" : 0
  }, {
    "time" : 1768924800000,
    "count" : 0
  }, {
    "time" : 1769011200000,
    "count" : 0
  }, {
    "time" : 1769097600000,
    "count" : 0
  }, {
    "time" : 1769184000000,
    "count" : 0
  }, {
    "time" : 1769270400000,
    "count" : 0
  }, {
    "time" : 1769356800000,
    "count" : 0
  }, {
    "time" : 1769443200000,
    "count" : 0
  }, {
    "time" : 1769529600000,
    "count" : 0
  } ]
 } ],
 "list_5xx" : [ {
  "error_code" : "500",
  "list" : [ ]
 } ]
}
```


Status code: 400

Failure response


```
{
 "error_msg" : "The project ID in the request does not match that in the token.",
 "error_code" : "ModelArts.0210"
}
```


Status Codes

Status Code	Description
200	Success response

Status Code	Description
400	Failure response

Error Codes

For details, see **Error Codes**.

6 Appendixes

6.1 Status Codes

The following table describes the status codes.

Table 6-1 Status codes

Status Code	Code	Status Code Description
100	Continue	The client should proceed with the request. This temporary response is used to inform the client that some requests have been received and not rejected by the server.
101	Switching Protocols	The protocol should be switched. The protocol can only be switched to a later version. For example, switch the current HTTP protocol to the latest version.
200	OK	The request has been fulfilled.
201	Created	The request for creating a resource has been fulfilled.
202	Accepted	The request has been accepted but is still being processed.
203	Non-Authoritative Information	Non-authoritative information. The request is successful.

Status Code	Code	Status Code Description
204	NoContent	The request has been fulfilled, but the HTTP response does not contain a response body. The status code is returned in response to an HTTP OPTIONS request.
205	Reset Content	The server has fulfilled the request, but the client is required to reset the content.
206	Partial Content	The server has processed a part of the GET request.
300	Multiple Choices	There are multiple options for the requested resource. The response contains a list of resource characteristics and addresses from which the user or user agent (such as a browser) can choose the most appropriate one.
301	Moved Permanently	The requested resource has been assigned a new permanent URI, and the new URI is contained in the response.
302	Found	The requested resource resides temporarily under a different URI.
303	See Other	The response to the request can be found under a different URI. It should be retrieved using a **GET** or **POST** method.
304	Not Modified	The requested resource has not been modified. When the server returns this status code, it does not return any resources.
305	Use Proxy	The requested resource must be accessed through a proxy.
306	Unused	The HTTP status code is no longer used.
400	BadRequest	The request is invalid. Modify the request and then try again.
401	Unauthorized	This status code is returned after the client provides the authentication information, indicating that the authentication information is incorrect or invalid.
402	Payment Required	This request is reserved.

Status Code	Code	Status Code Description
403	Forbidden	The request has been rejected. The server has received and understood the request; yet it refused to respond, because the request is set to deny access. Do not retry the request before modification.
404	NotFound	The requested resource cannot be found. Modify the request and then try again.
405	MethodNotAllowed	The request contains one or more methods not supported for the resource. Modify the request and then try again.
406	Not Acceptable	The server cannot implement the request based on the content characteristics of the request.
407	Proxy Authentication Required	This code is similar to 401, but indicates that the client must first authenticate itself with the proxy.
408	Request Time-out	The request timed out. The client may re-initiate the request without modifications at any time later.
409	Conflict	The request cannot be processed due to a conflict. This status code indicates that the resource that the client attempts to create already exits, or the requested update failed due to a conflict.
410	Gone	The requested resource is no longer available. The status code indicates that the requested resource has been deleted permanently.
411	Length Required	The server refuses to process the request without a defined **Content-Length**.
412	Precondition Failed	The server does not meet one of the requirements that the requester puts on the request.

Status Code	Code	Status Code Description
413	Request Entity Too Large	The request is larger than that a server can process. The server may disconnect the connection to prevent the client from continuously sending the request. If the server cannot process the request temporarily, the response will contain a **Retry-After** header field.
414	Request-URI Too Large	The request URI is too long for the server to process.
415	Unsupported Media Type	The media format in the request is not supported.
416	Requested range not satisfiable	The requested range is invalid.
417	Expectation Failed	The server fails to meet the requirements of the **Expect** request-header field.
422	UnprocessableEntity	The request is well-formed but is unable to process due to semantic errors.
429	TooManyRequests	The client sends too many requests to the server within a given time, exceeding the client's access frequency limit or beyond the server's processing capability. In this case, the client should retry after the time period specified in the **Retry-After** response header.
500	InternalServerError	The server is able to receive but unable to understand the request.
501	Not Implemented	The server does not support the requested function, and therefore cannot implement the request.
502	Bad Gateway	The server acting as a gateway or proxy has received an invalid request from a remote server.
503	ServiceUnavailable	The requested service is invalid. Modify the request and then try again.
504	ServerTimeout	The server cannot return a timely response. This status code is returned to the client only when the **Timeout** parameter is specified in the request.
505	HTTP Version not supported	The server does not support the HTTP protocol version used in the request.

6.2 Error Codes

The table below shows the error codes you might see when calling a MaaS API.

Table 6-2 Error codes

HTTP Status Code	Error Code	Error Message	Description	Handling
400	ModelArts.0104	CreateCustomEndpointReq.endpoint_name is a required field	The custom endpoint name is mandatory.	Check whether the custom endpoint name is specified.
400	ModelArts.6271	Custom endpoint name has already been used (including deleted access points). To ensure billing uniqueness, please use a different name.	The custom endpoint name has been used (including deleted endpoints). To ensure the uniqueness of the bill, use a different name.	Check whether the endpoint name has been used. If yes, use a different one.
400	ModelArts.6274	Custom endpoint name is invalid. Enter 1 to 64 characters, staring with a letter. Only letters, digits, point(.), underscores (_), and hyphens (-) are allowed	The name of the custom endpoint is invalid. Enter 1 to 64 characters. Only letters, digits, underscores (_), hyphens (-), and periods (.) are allowed.	Check the format of the custom endpoint name.
400	ModelArts.6291	CustomEndpointNotFound	The custom endpoint does not exist.	Check whether the custom endpoint has been created.
400	ModelArts.6297	Custom endpoint number exceed quota	The number of custom endpoints exceeds the quota.	Delete unnecessary custom endpoints and try again.

HTTP Status Code	Error Code	Error Message	Description	Handling
400	ModelArts.6299	Custom endpoint id is required	The custom endpoint ID is missing.	Check whether the custom endpoint ID is correct.
400	ModelArts.6613	Operation failed. The online service status needs to have the 'Mount to MAAS Gateway' switch turned on or the authentication method set to 'No Authentication' in order to create a custom custom endpoint.	Operation failed. You can create a custom endpoint only when MaaS gateway mounting is enabled or the authentication type is no authentication.	• In **Service Call Settings**, set **Authentication Mode** to **None**. • In **Network Settings**, enable **Public Network Access** and **Private Network Connection Approval**.
401	ModelArts.8503	No permission to create modelarts custom endpoints	You do not have the permission to create a custom endpoint for ModelArts real-time services.	Check whether the account permission meets the requirements.
400	ModelArts.2115	Invalid JSON request body	Invalid JSON request body.	The invalid request body has been modified based on the error message.
400	ModelArts.6298	Parameter (%s) security check failed, please check for special characters.	Security check for parameter (%s) failed. Check for special characters.	Check whether your parameters contain insecure characters.

6.3 Obtaining a Project ID and Name

Scenario

When you call an API, the project ID or name must be specified in some requests. The obtaining method is as follows:

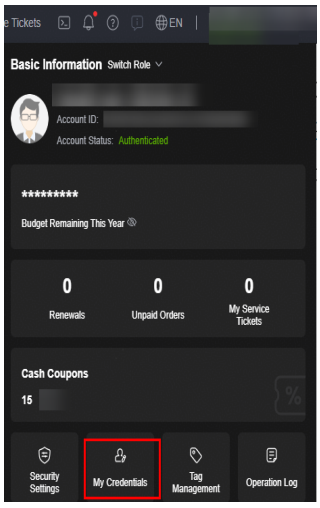
- **Obtaining a Project ID and Name from the Console**
- **Obtaining a Project ID by Calling an API**

Obtaining a Project ID and Name from the Console

To obtain the project ID (**project_id**) from the console, follow these steps:

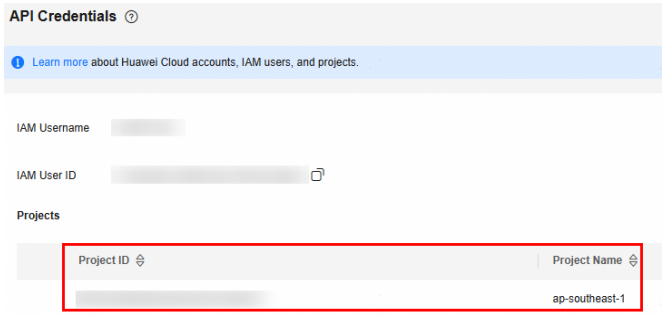
1. Log in to the **management console**.
2. Hover over the username in the upper-right corner and select **My Credentials** from the drop-down list.

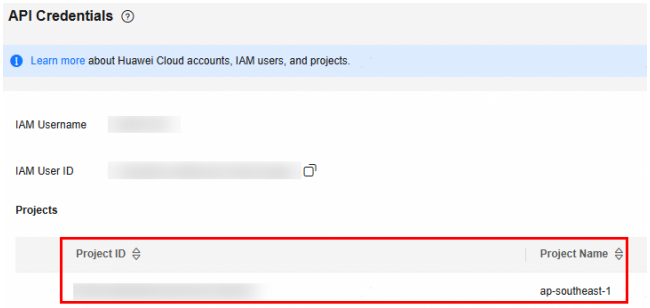
Figure 6-1 My Credentials



3. On the **API Credentials** page, the value in the **Project ID** column is the project ID, and the value in the **Project Name** column is the project name.

Figure 6-2 Viewing a project ID





If there are multiple projects in one region, expand **Region** and view subproject IDs in the **Project ID** column.

Obtaining a Project ID by Calling an API

A project ID can be obtained by calling a specific API. For details, see **Querying Project Information Based on the Specified Criteria**.

The API for obtaining a project ID is **GET https://*{iam-endpoint}*/v3/projects**. *{iam-endpoint}* can be obtained from **Regions and Endpoints**.

The following is an example response. For example, if MaaS is deployed in the **cn-southwest-2** region, the value of **name** in the response body is **cn-southwest-2**. The value of **id** under **projects** is the project ID.

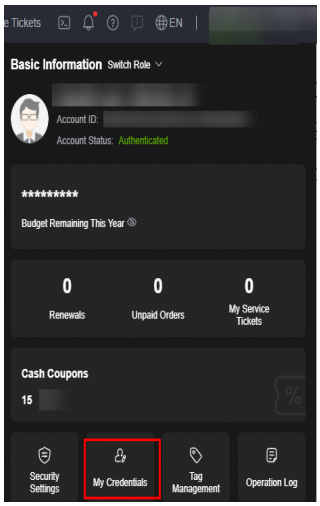
```
{
 "projects": [
  {
   "domain_id": "878991804cdc4ba597ae1403bdb******",
   "is_domain": false,
   "parent_id": "878991804cdc4ba597ae1403bd******",
   "name": "cn-southwest-2",
   "description": "",
   "links": {
    "next": null,
    "previous": null,
    "self": "https://iam.cn-southwest-2.myhuaweicloud.com/v3/projects/8d6c92e4e2d14ae685a697adfe******"
   },
   "id": "8d6c92e4e2d14ae685a697adfe******",
   "enabled": true
  },
 ],
 "links": {
  "next": null,
  "previous": null,
  "self": "https://iam.cn-southwest-2.myhuaweicloud.com/v3/projects?
domain_id=878991804cdc4ba597ae1403bd******"
 }
}
```

6.4 Obtaining an Account Name and Account ID

When you call APIs, certain requests require the account name and ID. To obtain an account name and ID, do as follows:

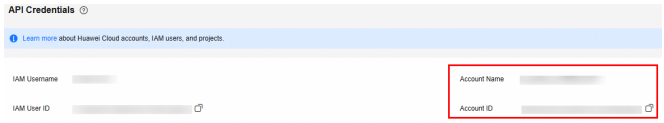
1. Log in to the **management console**.
2. Hover over the username in the upper-right corner and select **My Credentials** from the drop-down list.

Figure 6-3 My Credentials



3. On the **API Credentials** page, view the **Account Name** and **Account ID**.

Figure 6-4 Viewing the account name and ID

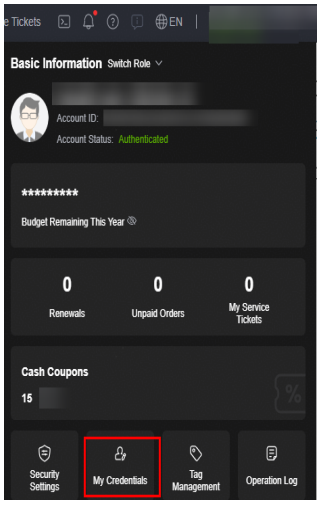


6.5 Obtaining a Username and ID

When you call APIs, certain requests require a username and ID. To obtain an account name and ID, do as follows:

1. Log in to the **management console**.
2. Hover over the username in the upper-right corner and select **My Credentials** from the drop-down list.

Figure 6-5 My Credentials



3. On the **API Credentials** page, obtain the IAM username and ID.

Figure 6-6 Viewing the username and ID