



PUBLIC  
2021-02-11

# Application Autoscaler

# Content

|          |  |          |
|----------|--|----------|
| <b>1</b> | <b>What Is Application Autoscaler?</b>             | <b>3</b> |
| <b>2</b> | <b>Initial Setup.</b>                              | <b>5</b> |
| <b>3</b> | <b>Development.</b>                                | <b>7</b> |
| 3.1      | Working with the Application Autoscaler Dashboard. | 7        |
|          | Access the Application Autoscaler Dashboard.       | 7        |
|          | Use the Application Autoscaler Dashboard.          | 8        |
| 3.2      | Defining a Scaling Policy.                         | 9        |
|          | Dynamic Scaling Policy.                            | 9        |
|          | Schedule-Based Scaling Policy.                     | 17       |
| 3.3      | Defining a Custom Metric.                          | 31       |
|          | Custom Metric API.                                 | 33       |

# 1 What Is Application Autoscaler?

Automatically scale your applications to meet their dynamic resource needs.

Application Autoscaler lets you automatically increase or decrease the number of your application instances based on the policies you have defined.

## Environment

This service runs in the Cloud Foundry environment.

## Features

|  |   |
|--|---|
| <b>Scale your applications automatically</b> | Automatically increase or decrease the number of application instances to meet the changing demands of application resources. |
| <b>Make your scaling resource-specific</b>   | Scale your applications on the basis of any standard metric type, such as memory consumed, CPU, response time, or throughput. |
| <b>Define custom metrics</b>                 | Scale your applications by defining custom metrics that match their requirements.   |
| <b>Manage your costs effectively</b>         | Only consume and pay for the resources, which your applications require.  |

## Regional Availability

Application Autoscaler is available in the following regions:

Regional Availability of Application Autoscaler

| IaaS Provider   | Region Name   |
|-----------------|---|
| AWS             | Japan (Tokyo), Canada (Montreal), Australia (Sydney), Brazil (São Paulo), US East (VA), Europe (Frankfurt), Singapore |
| Microsoft Azure | Singapore, Japan (Tokyo), US East (VA), US West (WA), Europe (Netherlands)  |
| GCP             | US Central (IA)   |

For more information, see [Regions](#).

## Trial

Trial accounts let you try out SAP Cloud Platform for free with a restricted use of the platform resources and services. For more information, see [Trial Accounts](#).

Application Autoscaler is available for trial use in the following regions:

Trial Availability of Application Autoscaler

| IaaS Provider   | Region Name   |
|-----------------|---|
| AWS             | Japan (Tokyo), Canada (Montreal), Australia (Sydney), Brazil (São Paulo), US East (VA), Europe (Frankfurt), Singapore |
| Microsoft Azure | Singapore, Japan (Tokyo), US East (VA), US West (WA), Europe (Netherlands)  |
| GCP             | US Central (IA)   |

## 2 Initial Setup

Create an instance of the Application Autoscaler service and bind it to your application.

### Prerequisites

- You have downloaded and set up the Cloud Foundry Command Line Interface (cf CLI). See [Download and Install the Cloud Foundry Command Line Interface](#).
- You are logged on to your Cloud Foundry space. See [Log On to the Cloud Foundry Environment Using the Cloud Foundry Command Line Interface](#).
- You are a Space Developer in the chosen Cloud Foundry space.
- You have deployed the application you want to scale in the Cloud Foundry environment. See [Deploy Business Applications in the Cloud Foundry Environment](#).

### Procedure

1. Check if the Application Autoscaler service is listed in the Service Marketplace using the following command:

```
cf marketplace
```

2. Create an instance of the service using the following command:

```
cf create-service autoscaler <service plan name> <instance name>
```

#### Sample Code

```
cf create-service autoscaler standard myservice
```

3. Bind the service instance to your application using the following command:

```
cf bind-service <application name> <instance name> -c <file name>.json
```

The JSON file contains the policy, which is needed to initiate the scaling of an application. For more information about how to define a policy, see [Defining a Scaling Policy \[page 9\]](#).

#### Note

If you are building a multitarget application (MTA), you can leverage the MTA deployment descriptor to provide the scaling policy. See [Service Binding Parameters](#).

4. Check if the instance is successfully bound to the application using the following command:

```
cf service <instance name>
```

The *Bound apps* field displays all applications bound to the instance.

## 3 Development

Use the Application Autoscaler in your development.

Using the Application Autoscaler in your development comprises the following steps:

- [Working with the Application Autoscaler Dashboard \[page 7\]](#)  
Access and use the Application Autoscaler dashboard.
- [Defining a Scaling Policy \[page 9\]](#)  
Define a policy to scale your application instances either dynamically or based on schedules.
- [Defining a Custom Metric \[page 31\]](#)  
Define your own metrics to scale applications based on your requirements.

### 3.1 Working with the Application Autoscaler Dashboard

Access and use the Application Autoscaler dashboard.

#### 3.1.1 Access the Application Autoscaler Dashboard

Access the Application Autoscaler dashboard, which lets you create or update a scaling policy, and view the scaling details.

#### Prerequisites

You have created an Application Autoscaler service instance and bound it to your application. See [Initial Setup \[page 5\]](#).

#### Procedure

1. In the SAP BTP cockpit, navigate to the **Organization** > **Space** containing the Application Autoscaler.
2. From the navigation pane, choose [Service Marketplace](#).
3. Choose [Application Autoscaler](#).
4. In the navigation pane, choose [Instances](#).

The list of created instances appears.

5. From the [Actions](#) column, choose (Open Dashboard).
6. Provide the platform credentials to access the dashboard.

The [Overview](#) page appears.

7. In the navigation pane, choose [Referencing Apps](#).

The list of applications bound to the service instance appears. In the [Actions](#) column, the [Manage Policy](#) and [Scaling History](#) options are available.

## 3.1.2 Use the Application Autoscaler Dashboard

Manage your scaling policy and view the scaling history.

The Application Autoscaler dashboard comprises two features, which let you perform different actions: [Manage Policy](#) and [Scaling History](#).

### Manage Policy

A scaling policy defines the scaling requirements of an application. [Manage Policy](#) lets you perform the following tasks:

- [Create a Policy](#)  
Define scaling requirements if you haven't used configuration parameters with the bind request during the binding of the service with an application.
- [View a Policy](#)  
Preview your policy, especially if you've defined it during the binding process.
- [Edit a Policy](#)  
Change an existing policy, for example to:
  - Add more scaling rules
  - Add schedule-based scaling along with dynamic scaling
  - Update scaling parameters like cool down and threshold adjustment based on their behavior as observed in the [Scaling History](#) feature.

For information about scaling policy, see [Dynamic Scaling Policy \[page 9\]](#).

### Scaling History

This feature lists all scaling events that were triggered by the Application Autoscaler service instance after it was bound to an application. The scaling history helps you to troubleshoot and resolve issues if the scaling doesn't match the defined policy.

#### ❖ Example

If the scaling failed due to quota restrictions, an appropriate message is displayed in the scaling history. Accordingly, you can request additional quota.



## 3.2 Defining a Scaling Policy

Define a policy to scale your application instances either dynamically or based on schedules.

To initiate the scaling of an application, you need to define a policy. A policy is a JSON file containing an array of rules or a single rule for scaling. This JSON file is passed while binding the service with the application to be scaled. For more information, see [Initial Setup \[page 5\]](#).

You can choose between the following types of scaling policies:

- [Dynamic Scaling Policy \[page 9\]](#)  
Scale your application instances based on memory/CPU usage, response time, throughput, or custom metrics.
- [Schedule-Based Scaling Policy \[page 17\]](#)  
Scale your application instances based on schedules.

### Related Information

[Parameters for a Dynamic Scaling Policy \[page 12\]](#)

[Parameters for a Schedule-Based Scaling Policy \[page 20\]](#)

[Time Zones for a Schedule-Based Policy \[page 22\]](#)

### 3.2.1 Dynamic Scaling Policy

Scale your application instances based on memory or CPU usage, response time, throughput, or custom metrics.

#### Metric Types memory consumed and memory utilized

These scaling rules define that if the memory consumption either exceeds or falls below the set threshold, the application is respectively scaled up or down by an instance. If an array of scaling rules is defined, the Application Autoscaler runs periodic checks of the scaling policy to verify if any of the specified conditions matches. New instances are generated based on specified rules.

#### ❖ Example

A scaling policy comprises two rules:

- Rule 1: If the memory exceeds 500 MB, a single instance of an application is created.
- Rule 2: If the memory exceeds 800 MB, two application instances are created.

When defining an array of rules, its sequence must match the descending order of the threshold values, so that the higher value appears first in the sequence. In this example, rule 2 must be defined before rule 1 to ensure that the Application Autoscaler checks all conditions.

The following table provides examples for dynamic scaling policies based on the metric types **memory consumed** and **memory utilized**.

| Metric Type memory consumed   | Metric Type memory utilized   |
|---|---|
| <p>≡ Sample Code</p> <pre> {   "instance_min_count": 1,   "instance_max_count": 5,   "scaling_rules": [     {       "metric_type": "memoryused",       "threshold": 90,       "operator": "&gt;=",       "adjustment": "+1"     },     {       "metric_type": "memoryused",       "threshold": 30,       "operator": "&lt;",       "adjustment": "-1"     }   ] }</pre> | <p>≡ Sample Code</p> <pre> {   "instance_min_count": 1,   "instance_max_count": 5,   "scaling_rules": [     {       "metric_type": "memoryutil",       "threshold": 90,       "operator": "&gt;=",       "adjustment": "+1"     },     {       "metric_type": "memoryutil",       "threshold": 30,       "operator": "&lt;",       "adjustment": "-1"     }   ] }</pre> |

## Metric Types throughput and responsetime

The following samples are simple scaling policies for the metric types **throughput** and **responsetime**. The scaling rule determines if the throughput or response time exceeds or falls below the set threshold.

## Sample Code

```
{
  "instance_min_count": 1,
  "instance_max_count": 5,
  "scaling_rules": [
    {
      "metric_type":
"throughput",
      "threshold": 100, //
the throughput value in Requests/
second
      "operator": ">",
      "adjustment": "+1"
    },
    {
      "metric_type":
"throughput",
      "threshold": 100,
      "operator": "<=",
      "adjustment": "-1"
    }
  ]
}
```

## Sample Code

```
{
  "instance_min_count": 1,
  "instance_max_count": 5,
  "scaling_rules": [
    {
      "metric_type":
"responsetime",
      "threshold": 900, //
the response time value in
milliseconds
      "operator": ">",
      "adjustment": "+1"
    },
    {
      "metric_type":
"responsetime",
      "threshold": 900,
      "operator": "<=",
      "adjustment": "-1"
    }
  ]
}
```

## Advanced Usage

For advanced usage, you can provide optional values along with the mandatory ones. The following sample displays such a scaling policy:

## Sample Code

```
{
  "instance_min_count": 1,
  "instance_max_count": 5,
  "scaling_rules": [
    {
      "metric_type": "memoryused",
      "breach_duration_secs": 600,
      "threshold": 90,
      "operator": ">=",
      "cool_down_secs": 300,
      "adjustment": "+1"
    },
    {
      "metric_type": "memoryused",
      "breach_duration_secs": 600,
      "threshold": 30,
      "operator": "<",
      "cool_down_secs": 300,
      "adjustment": "-1"
    }
  ]
}
```

For more information about the parameters used, see [Parameters for a Dynamic Scaling Policy \[page 12\]](#).

## Related Information

[Parameters for a Dynamic Scaling Policy \[page 12\]](#)

### 3.2.1.1 Parameters for a Dynamic Scaling Policy

Get to know the parameters used for dynamic scaling.

| Parameter          | Description   | Mandatory | Data Type | Value Range   | Default Value | Example |
|--------------------|---|-----------|-----------|---|---------------|---------|
| instance_min_count | The minimum number of application instances that are always running.  | Yes       | number    | minimum = 1; maximum = no upper limit   | none          | 1       |
| instance_max_count | The maximum number of application instances that can be provisioned as part of application scaling.   | Yes       | number    | minimum = at least one more than instance_min_count; maximum = no upper limit | none          | 5       |
| scaling_rules      | A rule comprises a group of key values set to automatically trigger the scaling activity. You can have one or more rules. At least one rule out of a possible array size of two is validated. | Yes       | array     | none  | none          | none    |

| Parameter                | Description   | Mandatory | Data Type | Value Range | Default Value | Example  |
|--------------------------|---|-----------|-----------|-------------|---------------|--|
| <code>metric_type</code> | The metric type can be memory usage in mebibytes ( <code>memoryused</code> ), memory utilization as a percentage of the memory quota ( <code>memoryutil</code> ), CPU as a percentage of virtual CPUs used ( <code>cpu</code> ), throughput in requests per second ( <code>throughput</code> ), or response time in milliseconds ( <code>responsetime</code> ). | Yes       | string    | none        | none          | One of the following options: <ul style="list-style-type: none"> <li><code>memoryused</code></li> <li><code>memoryutil</code></li> <li><code>cpu</code></li> <li><code>throughput</code></li> <li><code>responsetime</code></li> </ul> |

| Parameter            | Description   | Mandatory | Data Type | Value Range                                  | Default Value | Example     |
|----------------------|---|-----------|-----------|--|---------------|-------------|
| breach_duration_secs | <p>The duration to analyze the collected metrics data points. The duration is considered from the current time to the past, for example: the last 600 seconds.</p> <p>The service performs this analysis to check if the data points are consistently above or below the set threshold and makes the decision to scale. For a more accurate decision, set a larger duration, as in this case, more data points are analyzed.</p> <p>In situations that cause a rapid increase of application load in short intervals, set a small duration. This enables the application to scale out before the maximum memory threshold is exceeded and prevents the application from crashing.</p> | No        | number    | minimum = 60 seconds; maximum = 3600 seconds | 300 seconds   | 300 seconds |

| Parameter | Description  | Mandatory | Data Type | Value Range   | Default Value | Example |
|-----------|--|-----------|-----------|---|---------------|---------|
| threshold | The resources that can be considered for triggering the scaling activity are memory used in MB, memory utilized as percentage of memory assigned to the application, response time in milliseconds, or throughput in RPS. If any of the considered resource values is above or below the respective set values, the Application Autoscaler initiates the scaling activity. | Yes       | integer   | For memory consumed, minimum = 1; maximum = no upper limit.<br><br>For memory utilized, minimum = 1; maximum = 100<br><br>For throughput, minimum = 1; maximum = no upper limit.<br><br>For request per second, minimum = 1; maximum = no upper limit.<br><br>For cpu, minimum = 1; maximum = 100 | none          | 30      |
| operator  | The operator is used in combination with the threshold value to compare the current metric value. The expression is as follows:<br>[MetricUsage value]<br>[operator]<br>[threshold].   | Yes       | string    | none  | none          | 64 > 30 |

| Parameter      | Description  | Mandatory | Data Type | Value Range                                  | Default Value | Example   |
|----------------|--|-----------|-----------|--|---------------|---|
| cool_down_secs | The minimum duration between two successive scaling triggers. After the first scaling trigger, the second trigger occurs after the specified interval. This duration enables the application to be stable enough before it starts the next trigger. For production environments, a longer duration is recommended. | No        | number    | minimum = 60 seconds; maximum = 3600 seconds | 300 seconds   | 300   |
| adjustment     | The number of application instances to be scaled as part of the scaling activity. The format is + or – followed by the number of required instances.   | Yes       | number    | none   | none          | +1 to scale up by an instance; –2 to scale down by two instances. |

## Related Information

[Dynamic Scaling Policy \[page 9\]](#)



## 3.2.2 Schedule-Based Scaling Policy

Scale your application instances based on schedules.

You can either define recurring schedules or specific date schedules for scaling. If you want to define multiple schedules, make sure that they don't overlap.

### ⚠ Caution

Overlapping schedules don't trigger the scaling.

In recurring schedules, the date is optional. However, you need to provide a date for a specific date schedule. If you set the parameters `instance_min_count` and `instance_max_count` within a schedule, they override the global values set for these parameters.

## Recurring Schedules

The following example shows a schedule-based scaling policy:

### Sample Code

```
{
  "instance_min_count": 1,
  "instance_max_count": 5,
  "schedules": {
    "timezone": "Asia/Shanghai",
    "recurring_schedule": [
      {
        "start_time": "10:00",
        "end_time": "18:00",
        "days_of_week": [
          1,
          2,
          3
        ],
        "instance_min_count": 1,
        "instance_max_count": 10,
        "initial_min_instance_count": 5
      }
    ]
  }
}
```

The policy in the example defines a recurring schedule, which is set for the Asia/Shanghai time zone. The schedule triggers application instances during the specified time frame on Sunday, Monday, and Tuesday. The days of the week are specified through the numbers 1 to 7 starting with 1 for Monday.

The policy defines the minimum and maximum number of application instances during the time frame as well as the number of application instances that should be available at the beginning of the recurring schedule.

For more information about the time zones for schedule-based policies, see [Time Zones for a Schedule-Based Policy \[page 22\]](#).

## Schedules with Start Day and Day of the Month, and Specific Date Schedules

The following examples show schedule-based scaling with start date and day of the month, as well as a specific date schedule.

### Schedule with Start Date and Day of the Month

#### Sample Code

```
"schedules": {
  "timezone": "Asia/Shanghai",
  "recurring_schedule": [
    {
      "start_date": "2016-06-27",
      "end_date": "2016-07-23",
      "start_time": "11:00",
      "end_time": "19:30",
      "days_of_month": [
        5,
        15,
        25
      ],
      "instance_min_count": 3,
      "instance_max_count": 10,
      "initial_min_instance_count": 5
    }
  ]
}
```

### Specific Date Schedule

#### Sample Code

```
{
  "instance_min_count": 1,
  "instance_max_count": 5,
  "schedules": {
    "timezone": "Asia/Shanghai",
    "specific_date": [
      {
        "start_date_time": "2015-06-02T10:00",
        "end_date_time": "2015-06-15T13:59",
        "instance_min_count": 1,
        "instance_max_count": 4,
        "initial_min_instance_count": 2
      },
      {
        "start_date_time": "2015-01-04T20:00",
        "end_date_time": "2015-02-19T23:15",
        "instance_min_count": 2,
        "instance_max_count": 5,
        "initial_min_instance_count": 3
      }
    ]
  }
}
```

## Advanced Usage

For advanced usage, you can provide optional values along with the mandatory ones. The following sample displays such a scaling policy:

#### Sample Code

```
{
  "instance_min_count": 1,
  "instance_max_count": 5,
  "schedules": {
```

```

"timezone": "Asia/Shanghai",
"recurring_schedule": [
  {
    "start_time": "10:00",
    "end_time": "18:00",
    "days_of_week": [
      1,
      2,
      3
    ],
    "instance_min_count": 1,
    "instance_max_count": 10,
    "initial_min_instance_count": 5
  },
  {
    "start_date": "2016-06-27",
    "end_date": "2016-07-23",
    "start_time": "11:00",
    "end_time": "19:30",
    "days_of_month": [
      5,
      15,
      25
    ],
    "instance_min_count": 3,
    "instance_max_count": 10,
    "initial_min_instance_count": 5
  }
]
}

```

For more information about the parameters used, see [Parameters for a Schedule-Based Scaling Policy \[page 20\]](#).

## Related Information

[Parameters for a Schedule-Based Scaling Policy \[page 20\]](#)

[Time Zones for a Schedule-Based Policy \[page 22\]](#)

### 3.2.2.1 Parameters for a Schedule-Based Scaling Policy

Get to know the parameters used for recurring schedules or specific date schedules.

| Parameter          | Description  | Mandatory | Data Type | Value Range                      | Default Value | Example      |
|--------------------|--|-----------|-----------|----------------------------------|---------------|--------------|
| schedules          | A schedule lets you configure scaling rules for specific days or on a recurring basis. Schedules guard against expected high surges or low activity periods. | Yes       | none      | none                             | none          | none         |
| timezone           | A valid time zone to run the schedule. For more information, see <a href="#">Time Zones for a Schedule-Based Policy [page 22]</a> .                          | Yes       | string    | none                             | none          | Asia/Sanghai |
| recurring_schedule | Triggers the scaling rule recursively during the specified intervals   | Yes       | none      | none                             | none          | none         |
| start_time         | Start time of a recurring schedule in 24-hr format (HH:MM).  | Yes       | string    | minimum = 00:00; maximum = 23:59 | none          | 10:00        |
| end_time           | End time of a recurring schedule in 24-hr format (HH:MM).  | Yes       | string    | minimum = 00:00; maximum = 23:59 | none          | 21:00        |

| Parameter          | Description  | Mandatory | Data Type | Value Range   | Default Value | Example         |
|--------------------|--|-----------|-----------|---|---------------|-----------------|
| start_date         | Start date of the schedule in YYYY-MM-DD format.   | No        | string    | minimum = current date; maximum = no upper limit              | none          | 2020-06-27      |
| end_date           | End date of the schedule in YYYY-MM-DD format.   | No        | string    | minimum = set by start_date; maximum = no upper limit         | none          | 2020-07-23      |
| days_of_week       | Triggers the scaling on weekdays ranging from 1 (Monday) to 7 (Sunday). The rule is executed during the weekdays specified within the array. | Yes       | array     | minimum = 1; maximum = 7                                      | none          | [1, 3, 5]       |
| days_of_month      | Triggers the scaling on days of the month ranging from 1 to 31. The rule is executed during the specified days of the month.                 | Yes       | array     | minimum = 1; maximum = 31                                     | none          | [1, 11, 24, 30] |
| instance_min_count | Minimum number of instances during the recurrence period.  | Yes       | number    | minimum = 1; maximum = no upper limit                         | none          | 1               |
| instance_max_count | Maximum number of instances during the recurrence period.  | Yes       | number    | minimum = set by instance_min_count; maximum = no upper limit | none          | 5               |

| Parameter                      | Description  | Mandatory | Data Type | Value Range   | Default Value | Example          |
|--------------------------------|--|-----------|-----------|---|---------------|------------------|
| initial_minimum_instance_count | Minimum number of instances to scale up at the beginning of the recurrence period. | Yes       | number    | minimum = set by instance_minimum_count; maximum = instance_maximum_count | none          | 5                |
| specific_date                  | Triggers the scaling rule during the specified date intervals.                     | Yes       | none      | none  | none          | none             |
| start_date_time                | Start date and time of the schedule in YYYY-MM-DDTHH:MM format                     | Yes       | string    | minimum = current date and time; maximum = no upper limit                 | none          | 2015-06-02T10:00 |
| end_date_time                  | End date and time of the schedule in YYYY-MM-DDTHH:MM format                       | Yes       | string    | minimum = set by start_date_time; maximum = no upper limit                | none          | 2015-09-02T10:00 |

## Related Information

[Schedule-Based Scaling Policy \[page 17\]](#)

[Time Zones for a Schedule-Based Policy \[page 22\]](#)

### 3.2.2.2 Time Zones for a Schedule-Based Policy

Get to know the time zones supported for recurring schedules or specific date schedules.

The following code block lists all time zones supported by the Application Autoscaler.

```
{
  "Etc/GMT+12",
  "Etc/GMT+11",
  "Pacific/Midway",
  "Pacific/Niue",
}
```

```

"Pacific/Pago_Pago",
"Pacific/Samoa",
"US/Samoa",
"Etc/GMT+10",
"HST",
"Pacific/Honolulu",
"Pacific/Johnston",
"Pacific/Rarotonga",
"Pacific/Tahiti",
"US/Hawaii",
"Pacific/Marquesas",
"America/Adak",
"America/Atka",
"Etc/GMT+9",
"Pacific/Gambier",
"US/Aleutian",
"America/Anchorage",
"America/Juneau",
"America/Metlakatla",
"America/Nome",
"America/Sitka",
"America/Yakutat",
"Etc/GMT+8",
"Pacific/Pitcairn",
"US/Alaska",
"America/Creston",
"America/Dawson",
"America/Dawson_Creek",
"America/Ensenada",
"America/Hermosillo",
"America/Los_Angeles",
"America/Phoenix",
"America/Santa_Isabel",
"America/Tijuana",
"America/Vancouver",
"America/Whitehorse",
"Canada/Pacific",
"Canada/Yukon",
"Etc/GMT+7",
"MST",
"Mexico/BajaNorte",
"PST8PDT",
"US/Arizona",
"US/Pacific",
"US/Pacific-New",
"America/Belize",
"America/Boise",
"America/Cambridge_Bay",
"America/Chihuahua",
"America/Costa_Rica",
"America/Denver",
"America/Edmonton",
"America/El_Salvador",
"America/Guatemala",
"America/Inuvik",
"America/Managua",
"America/Mazatlan",
"America/Ojinaga",
"America/Regina",
"America/Shiprock",
"America/Swift_Current",
"America/Tegucigalpa",
"America/Yellowknife",
"Canada/East-Saskatchewan",
"Canada/Mountain",
"Canada/Saskatchewan",
"Etc/GMT+6",
"MST7MDT",

```

```
"Mexico/BajaSur",
"Navajo",
"Pacific/Galapagos",
"US/Mountain",
"America/Atikokan",
"America/Bahia_Banderas",
"America/Bogota",
"America/Cancun",
"America/Cayman",
"America/Chicago",
"America/Coral_Harbour",
"America/Eirunepe",
"America/Guayaquil",
"America/Indiana/Knox",
"America/Indiana/Tell_City",
"America/Jamaica",
"America/Knox_IN",
"America/Lima",
"America/Matamoros",
"America/Menominee",
"America/Merida",
"America/Mexico_City",
"America/Monterrey",
"America/North_Dakota/Beulah",
"America/North_Dakota/Center",
"America/North_Dakota/New_Salem",
"America/Panama",
"America/Porto_Acre",
"America/Rainy_River",
"America/Rankin_Inlet",
"America/Resolute",
"America/Rio_Branco",
"America/Winnipeg",
"Brazil/Acre",
"CST6CDT",
"Canada/Central",
"Chile/EasterIsland",
"EST",
"Etc/GMT+5",
"Jamaica",
"Mexico/General",
"Pacific/Easter",
"US/Central",
"US/Indiana-Starke",
"America/Caracas",
"America/Anguilla",
"America/Antigua",
"America/Aruba",
"America/Asuncion",
"America/Barbados",
"America/Blanc-Sablon",
"America/Boa_Vista",
"America/Campo_Grande",
"America/Cuiaba",
"America/Curacao",
"America/Detroit",
"America/Dominica",
"America/Fort_Wayne",
"America/Grand_Turk",
"America/Grenada",
"America/Guadeloupe",
"America/Guyana",
"America/Havana",
"America/Indiana/Indianapolis",
"America/Indiana/Marengo",
"America/Indiana/Petersburg",
"America/Indiana/Vevay",
"America/Indiana/Vincennes",
```



```

"America/Indiana/Winamac",
"America/Indianapolis",
"America/Iqaluit ",
"America/Kentucky/Louisville ",
"America/Kentucky/Monticello",
"America/Kralendijk",
"America/La_Paz",
"America/Louisville ",
"America/Lower_Princes",
"America/Manaus",
"America/Marigot",
"America/Martinique",
"America/Montreal",
"America/Montserrat",
"America/Nassau",
"America/New_York",
"America/Nipigon",
"America/Pangnirtung ",
"America/Port-au-Prince ",
"America/Port_of_Spain",
"America/Porto_Velho",
"America/Puerto_Rico ",
"America/Santo_Domingo ",
"America/St_Barthelemy",
"America/St_Kitts",
"America/St_Lucia",
"America/St_Thomas",
"America/St_Vincent",
"America/Thunder_Bay",
"America/Toronto",
"America/Tortola",
"America/Virgin",
"Brazil/West",
"Canada/Eastern",
"Cuba",
"EST5EDT",
"Etc/GMT+4",
"US/East-Indiana",
"US/Eastern",
"US/Michigan",
"America/Araguaina ",
"America/Argentina/Buenos_Aires ",
"America/Argentina/Catamarca ",
"America/Argentina/ComodRivadavia ",
"America/Argentina/Cordoba ",
"America/Argentina/Jujuy ",
"America/Argentina/La_Rioja ",
"America/Argentina/Mendoza ",
"America/Argentina/Rio_Gallegos ",
"America/Argentina/Salta ",
"America/Argentina/San_Juan ",
"America/Argentina/San_Luis ",
"America/Argentina/Tucuman ",
"America/Argentina/Ushuaia",
"America/Bahia",
"America/Belem",
"America/Buenos_Aires",
"America/Catamarca",
"America/Cayenne",
"America/Cordoba",
"America/Fortaleza",
"America/Glace_Bay",
"America/Goose_Bay",
"America/Halifax",
"America/Jujuy",
"America/Maceio",
"America/Mendoza",
"America/Moncton",

```

```

"America/Montevideo",
"America/Paramaribo",
"America/Recife",
"America/Rosario",
"America/Santarem",
"America/Santiago",
"America/Sao_Paulo",
"America/Thule",
"Antarctica/Palmer",
"Antarctica/Rothera",
"Atlantic/Bermuda",
"Atlantic/Stanley",
"Brazil/East",
"Canada/Atlantic",
"Chile/Continental",
"Etc/GMT+3",
"America/St_Johns",
"Canada/Newfoundland",
"America/Godthab",
"America/Miquelon",
"America/Noronha ",
"Atlantic/South_Georgia",
"Brazil/DeNoronha",
"Etc/GMT+2",
"Atlantic/Cape_Verde",
"Etc/GMT+1",
"Africa/Abidjan",
"Africa/Accra",
"Africa/Bamako",
"Africa/Banjul",
"Africa/Bissau",
"Africa/Conakry",
"Africa/Dakar",
"Africa/Freetown",
"Africa/Lome",
"Africa/Monrovia",
"Africa/Nouakchott",
"Africa/Ouagadougou",
"Africa/Sao_Tome",
"Africa/Timbuktu",
"America/Danmarkshavn",
"America/Scoresbysund",
"Atlantic/Azores",
"Atlantic/Reykjavik",
"Atlantic/St_Helena",
"Etc/GMT",
"Etc/GMT+0",
"Etc/GMT-0",
"Etc/GMT0",
"Etc/Greenwich",
"Etc/UCT",
"Etc/UTC",
"Etc/Universal",
"Etc/Zulu",
"GMT",
"GMT+0",
"GMT-0",
"GMT0",
"Greenwich",
"Iceland",
"UCT",
"UTC",
"Universal",
"Zulu",
"Africa/Algiers",
"Africa/Bangui",
"Africa/Brazzaville",
"Africa/Casablanca",

```

```
"Africa/Douala",
"Africa/El_Aaiun",
"Africa/Kinshasa",
"Africa/Lagos",
"Africa/Libreville",
"Africa/Luanda",
"Africa/Malabo",
"Africa/Ndjamena",
"Africa/Niamey",
"Africa/Porto-Novo",
"Africa/Tunis",
"Africa/Windhoek",
"Atlantic/Canary",
"Atlantic/Faeroe",
"Atlantic/Faroe",
"Atlantic/Madeira",
"Eire",
"Etc/GMT-1",
"Europe/Belfast",
"Europe/Dublin",
"Europe/Guernsey",
"Europe/Isle_of_Man",
"Europe/Jersey",
"Europe/Lisbon",
"Europe/London",
"GB",
"GB-Eire",
"Portugal",
"WET",
"Africa/Blantyre",
"Africa/Bujumbura",
"Africa/Cairo",
"Africa/Ceuta",
"Africa/Gaborone",
"Africa/Harare",
"Africa/Johannesburg",
"Africa/Kigali",
"Africa/Lubumbashi",
"Africa/Lusaka",
"Africa/Maputo",
"Africa/Maseru",
"Africa/Mbabane",
"Africa/Tripoli",
"Antarctica/Troll",
"Arctic/Longyearbyen",
"Atlantic/Jan_Mayen",
"CET",
"Egypt",
"Etc/GMT-2",
"Europe/Amsterdam",
"Europe/Andorra",
"Europe/Belgrade",
"Europe/Berlin",
"Europe/Bratislava",
"Europe/Brussels",
"Europe/Budapest",
"Europe/Busingen",
"Europe/Copenhagen",
"Europe/Gibraltar",
"Europe/Kaliningrad",
"Europe/Ljubljana",
"Europe/Luxembourg",
"Europe/Madrid",
"Europe/Malta",
"Europe/Monaco",
"Europe/Oslo",
"Europe/Paris",
"Europe/Podgorica",
```

```
"Europe/Prague",
"Europe/Rome",
"Europe/San_Marino",
"Europe/Sarajevo",
"Europe/Skopje",
"Europe/Stockholm",
"Europe/Tirane",
"Europe/Vaduz",
"Europe/Vatican",
"Europe/Vienna",
"Europe/Warsaw",
"Europe/Zagreb",
"Europe/Zurich",
"Libya",
"MET",
"Poland",
"Africa/Addis_Ababa",
"Africa/Asmara",
"Africa/Asmera",
"Africa/Dar_es_Salaam",
"Africa/Djibouti",
"Africa/Juba",
"Africa/Kampala",
"Africa/Khartoum",
"Africa/Mogadishu",
"Africa/Nairobi",
"Antarctica/Syowa",
"Asia/Aden",
"Asia/Amman",
"Asia/Baghdad",
"Asia/Bahrain",
"Asia/Beirut",
"Asia/Damascus",
"Asia/Gaza",
"Asia/Hebron",
"Asia/Istanbul",
"Asia/Jerusalem",
"Asia/Kuwait",
"Asia/Nicosia",
"Asia/Qatar",
"Asia/Riyadh",
"Asia/Tel_Aviv",
"EET",
"Etc/GMT-3",
"Europe/Athens",
"Europe/Bucharest",
"Europe/Chisinau",
"Europe/Helsinki",
"Europe/Istanbul",
"Europe/Kiev",
"Europe/Mariehamn",
"Europe/Minsk",
"Europe/Moscow",
"Europe/Nicosia",
"Europe/Riga",
"Europe/Simferopol",
"Europe/Sofia",
"Europe/Tallinn",
"Europe/Tiraspol",
"Europe/Uzhgorod",
"Europe/Vilnius",
"Europe/Volgograd",
"Europe/Zaporozhye",
"Indian/Antananarivo",
"Indian/Comoro",
"Indian/Mayotte",
"Israel",
"Turkey",
```

"W-SU",  
"Asia/Dubai",  
"Asia/Muscat",  
"Asia/Tbilisi",  
"Asia/Yerevan",  
"Etc/GMT-4",  
"Europe/Samara",  
"Indian/Mahe",  
"Indian/Mauritius",  
"Indian/Reunion",  
"Asia/Kabul",  
"Asia/Tehran",  
"Iran",  
"Antarctica/Mawson",  
"Asia/Aqtau",  
"Asia/Aqtobe",  
"Asia/Ashgabat",  
"Asia/Ashkhabad",  
"Asia/Baku",  
"Asia/Dushanbe",  
"Asia/Karachi",  
"Asia/Oral",  
"Asia/Samarkand",  
"Asia/Tashkent",  
"Asia/Yekaterinburg",  
"Etc/GMT-5",  
"Indian/Kerguelen",  
"Indian/Maldives",  
"Asia/Calcutta",  
"Asia/Colombo",  
"Asia/Kolkata",  
"Asia/Kathmandu",  
"Asia/Katmandu",  
"Antarctica/Vostok",  
"Asia/Almaty",  
"Asia/Bishkek",  
"Asia/Dacca",  
"Asia/Dhaka",  
"Asia/Kashgar",  
"Asia/Novosibirsk",  
"Asia/Omsk",  
"Asia/Qyzylorda",  
"Asia/Thimbu",  
"Asia/Thimphu",  
"Asia/Urumqi",  
"Etc/GMT-6",  
"Indian/Chagos",  
"Asia/Rangoon",  
"Indian/Cocos",  
"Antarctica/Davis",  
"Asia/Bangkok",  
"Asia/Ho\_Chi\_Minh",  
"Asia/Hovd",  
"Asia/Jakarta",  
"Asia/Krasnoyarsk",  
"Asia/Novokuznetsk",  
"Asia/Phnom\_Penh",  
"Asia/Pontianak",  
"Asia/Saigon",  
"Asia/Vientiane",  
"Etc/GMT-7",  
"Indian/Christmas",  
"Antarctica/Casey",  
"Asia/Brunei",  
"Asia/Chita",  
"Asia/Choibalsan",  
"Asia/Chongqing",  
"Asia/Chungking",

```
"Asia/Harbin",
"Asia/Hong_Kong",
"Asia/Irkutsk",
"Asia/Kuala_Lumpur",
"Asia/Kuching",
"Asia/Macao",
"Asia/Macau",
"Asia/Makassar",
"Asia/Manila",
"Asia/Shanghai",
"Asia/Singapore",
"Asia/Taipei",
"Asia/Ujung_Pandang",
"Asia/Ulaanbaatar",
"Asia/Ulan_Bator",
"Australia/Perth",
"Australia/West",
"Etc/GMT-8",
"Hongkong",
"PRC",
"ROC",
"Singapore",
"Australia/Eucla",
"Asia/Dili",
"Asia/Jayapura",
"Asia/Khandyga",
"Asia/Pyongyang",
"Asia/Seoul",
"Asia/Tokyo",
"Asia/Yakutsk",
"Etc/GMT-9",
"Japan",
"Pacific/Palau",
"ROK",
"Australia/Adelaide ",
"Australia/Broken_Hill",
"Australia/Darwin",
"Australia/North",
"Australia/South",
"Australia/Yancowinna ",
"Antarctica/DumontDUrville",
"Asia/Magadan",
"Asia/Sakhalin",
"Asia/Ust-Nera",
"Asia/Vladivostok",
"Australia/ACT",
"Australia/Brisbane",
"Australia/Canberra",
"Australia/Currie",
"Australia/Hobart",
"Australia/Lindeman",
"Australia/Melbourne",
"Australia/NSW",
"Australia/Queensland",
"Australia/Sydney",
"Australia/Tasmania",
"Australia/Victoria",
"Etc/GMT-10",
"Pacific/Chuuk",
"Pacific/Guam",
"Pacific/Port_Moresby",
"Pacific/Saipan",
"Pacific/Truk",
"Pacific/Yap",
"Australia/LHI",
"Australia/Lord_Howe",
"Antarctica/Macquarie",
"Asia/Srednekolymsk",
```

```
"Etc/GMT-11",
"Pacific/Bougainville",
"Pacific/Efate",
"Pacific/Guadalcanal",
"Pacific/Kosrae",
"Pacific/Noumea",
"Pacific/Pohnpei",
"Pacific/Ponape",
"Pacific/Norfolk",
"Antarctica/McMurdo",
"Antarctica/South_Pole",
"Asia/Anadyr",
"Asia/Kamchatka",
"Etc/GMT-12",
"Kwajalein",
"NZ",
"Pacific/Auckland",
"Pacific/Fiji",
"Pacific/Funafuti",
"Pacific/Kwajalein",
"Pacific/Majuro",
"Pacific/Nauru",
"Pacific/Tarawa",
"Pacific/Wake",
"Pacific/Wallis",
"NZ-CHAT",
"Pacific/Chatham",
"Etc/GMT-13",
"Pacific/Apia",
"Pacific/Enderbury",
"Pacific/Fakaofo",
"Pacific/Tongatapu",
"Etc/GMT-14",
"Pacific/Kiritimati"
}
```

## Related Information

[Schedule-Based Scaling Policy \[page 17\]](#)

[Parameters for a Schedule-Based Scaling Policy \[page 20\]](#)

## 3.3 Defining a Custom Metric

Define your own metrics to scale applications based on your requirements.

### Context

As an alternative to the standard metrics provided by the Application Autoscaler, you can also define custom metrics such as memory consumed in megabytes or in percentage, response time, and throughput for scaling. Custom metrics offer more flexibility, so that you can scale applications based on your own requirements.

To use custom metrics, you need to perform the following tasks:

1. Report a custom metric to the Application Autoscaler by defining a policy.
2. Emit custom metric details using the HTTP API at regular intervals.

The following example shows a policy with a custom metric:

```
{
  "instance_min_count":1,
  "instance_max_count":4,
  "scaling_rules":[
    {
      "metric_type":"jobqueue",
      "breach_duration_secs":60,
      "threshold":100,
      "operator":>=",
      "cool_down_secs":120,
      "adjustment":"+1"
    }
  ]
}
```

### Note

The metric type used for custom metrics must not be any of the standard metric types.

### → Tip

We recommend a minimum duration of one minute between successive emissions of a custom metric.

## Procedure

1. Bind your application with the Application Autoscaler service instance using a policy that contains custom metrics scaling rule.
2. Read custom metrics binding credentials from the application environment.

### Sample Code

```
"custom_metrics": {
  "password": "73304df3-38a6-463a-880d-1451beedb49",
  "url": "https://metrics.bosh-lite.com",
  "username": "eb36e021-3441-4799-bf00-91ad6fcdd231"
}
```

3. Use the username and password for basic authentication for pushing custom metrics to the Application Autoscaler.
4. Push custom metrics at an interval of your choice using APIs. The REST API specification is the following:

```
POST /v1/apps/:<app_guid>/metrics
```

### Sample Code

```
POST /v1/apps/79d101b5-5379-45d8-8974-41c677ef34c9/metrics
Request Body
{
```



```

    "instance_index":0,
    "metrics":[
      {
        "name":"custom_metric1",
        "type":"gauge",
        "value":200,
        "unit":"unit"
      }
    ]
  }
}

```

## Related Information

[Custom Metric API \[page 33\]](#)

### 3.3.1 Custom Metric API

## Authentication

The credentials must be read from the environment of the application and passed in a basic authentication header to the API. You can retrieve these credentials from the `VCAP_SERVICES` environment variable. The `VCAP_SERVICES` variable looks as follows:

```

{
  "VCAP_SERVICES": {
    "autoscaler": [
      {
        "binding_name": null,
        "credentials": {
          "custom_metrics": {
            "password": [password],
            "url": [base url],
            "username": [username]
          }
        },
        "instance_name": [service_instance_name],
        "label": "autoscaler",
        "name": [service_name],
        "plan": [plan_name],
        "provider": null,
        "syslog_drain_url": null,
        "tags": [
          "autoscaler",
          "app-autoscaler",
          "cf-autoscaler"
        ],
        "volume_mounts": []
      }
    ]
  }
}

```

```
}
```

## API

The following example shows a sample request body for the metric:

```
POST /v1/apps/[app_guid]/metrics
Request Body
{
  "instance_index":0,
  "metrics":[
    {
      "name":"custom_metric1",
      "value":200,
      "unit":"unit"
    }
  ]
}
```

The following table explains the request body parameter details:



| Parameter           | Type           | Description  |
|---------------------|----------------|--|
| app_guid            | string         | GUID of the application.   |
| <b>Request Body</b> |                |  |
| instance_index      | integer        | The index of the application instance as reported by cloud controller. |
| metrics             | list of metric | The metrics that are to be emitted.                                    |
| <b>Metric</b>       |                |  |
| name                | string         | Name of the metric.  |
| value               | float64        | Value of the metric.   |
| unit                | string         | Value of the unit.   |

# Important Disclaimers and Legal Information

## Hyperlinks

Some links are classified by an icon and/or a mouseover text. These links provide additional information.

About the icons:

- Links with the icon  : You are entering a Web site that is not hosted by SAP. By using such links, you agree (unless expressly stated otherwise in your agreements with SAP) to this:
  - The content of the linked-to site is not SAP documentation. You may not infer any product claims against SAP based on this information.
  - SAP does not agree or disagree with the content on the linked-to site, nor does SAP warrant the availability and correctness. SAP shall not be liable for any damages caused by the use of such content unless damages have been caused by SAP's gross negligence or willful misconduct.
- Links with the icon  : You are leaving the documentation for that particular SAP product or service and are entering a SAP-hosted Web site. By using such links, you agree that (unless expressly stated otherwise in your agreements with SAP) you may not infer any product claims against SAP based on this information.

## Videos Hosted on External Platforms

Some videos may point to third-party video hosting platforms. SAP cannot guarantee the future availability of videos stored on these platforms. Furthermore, any advertisements or other content hosted on these platforms (for example, suggested videos or by navigating to other videos hosted on the same site), are not within the control or responsibility of SAP.

## Beta and Other Experimental Features

Experimental features are not part of the officially delivered scope that SAP guarantees for future releases. This means that experimental features may be changed by SAP at any time for any reason without notice. Experimental features are not for productive use. You may not demonstrate, test, examine, evaluate or otherwise use the experimental features in a live operating environment or with data that has not been sufficiently backed up.

The purpose of experimental features is to get feedback early on, allowing customers and partners to influence the future product accordingly. By providing your feedback (e.g. in the SAP Community), you accept that intellectual property rights of the contributions or derivative works shall remain the exclusive property of SAP.

## Example Code

Any software coding and/or code snippets are examples. They are not for productive use. The example code is only intended to better explain and visualize the syntax and phrasing rules. SAP does not warrant the correctness and completeness of the example code. SAP shall not be liable for errors or damages caused by the use of example code unless damages have been caused by SAP's gross negligence or willful misconduct.

## Gender-Related Language

We try not to use gender-specific word forms and formulations. As appropriate for context and readability, SAP may use masculine word forms to refer to all genders.

© 2021 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company. The information contained herein may be changed without prior notice.

Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies.

Please see <https://www.sap.com/about/legal/trademark.html> for additional trademark information and notices.