

PUBLIC  
2026-05-12

# SAP AI Core

# Content

- 1 What Is SAP AI Core? . . . . . 5**
- 1.1 Metering and Pricing for SAP AI Core. . . . . 8
  - Metering and Pricing for Generative AI. . . . . 9
  - Metering and Pricing for Predictive AI. . . . . 13
- 2 What's New for SAP AI Core . . . . . 17**
- 3 Concepts. . . . . 48**
- 3.1 AI API Overview. . . . . 55
  - AI API Runtime Implementations. . . . . 55
- 3.2 Resource Groups. . . . . 58
  - Scope of Resources. . . . . 59
- 3.3 Generative AI Hub in SAP AI Core Overview. . . . . 60
- 4 Initial Setup. . . . . 62**
- 4.1 Enabling the Service in Cloud Foundry. . . . . 62
  - Create a Subaccount. . . . . 63
  - Enable Cloud Foundry. . . . . 65
  - Create a Space. . . . . 66
  - Add a Service Plan. . . . . 67
  - Create a Service Instance. . . . . 73
  - Create a Service Key. . . . . 77
  - Use a Service Key. . . . . 80
- 4.2 Enabling the Service in the Kyma Environment. . . . . 83
- 5 Tutorials. . . . . 84**
- 6 Administration. . . . . 85**
- 6.1 Manage Your Git Repository. . . . . 85
  - Add a Git Repository. . . . . 85
  - Edit a Git Repository. . . . . 88
  - Delete a Git Repository. . . . . 89
- 6.2 Manage Applications. . . . . 91
  - Create an Application. . . . . 91
  - List Applications. . . . . 94
  - Edit an Application. . . . . 94
  - Delete an Application. . . . . 95
- 6.3 Manage Resource Groups. . . . . 95

	Resource Group Level Resources. . . . .	96
	Create a Resource Group. . . . .	97
	Edit a Resource Group. . . . .	99
	Delete a Resource Group. . . . .	100
6.4	Manage Object Store Secrets. . . . .	102
	Register an Object Store Secret. . . . .	102
	Edit an Object Store Secret. . . . .	107
	Delete an Object Store Secret. . . . .	112
6.5	Manage Docker Registry Secrets. . . . .	113
	Register Your Docker Registry Secret. . . . .	113
	Edit a Docker Registry Secret. . . . .	116
	Delete a Docker Registry Secret. . . . .	117
6.6	Manage Generic Secrets. . . . .	118
	Create a Generic Secret. . . . .	118
	Get Generic Secrets. . . . .	122
	Update a Generic Secret. . . . .	125
	Delete a Generic Secret. . . . .	128
	Consume Generic Secrets in Executions or Deployments. . . . .	129
6.7	Manage mTLS Certificate Secrets. . . . .	130
	Create an mTLS Certificate Secret. . . . .	132
	Consume mTLS Certificate Secrets. . . . .	135
	List mTLS Certificate Secrets. . . . .	137
	Get Details of an mTLS Certificate Secret . . . . .	138
	Rotate an mTLS Certificate Secret. . . . .	139
	Delete an mTLS certificate secret. . . . .	141
<b>7</b>	<b>APIs and API Extensions. . . . .</b>	<b>143</b>
<b>8</b>	<b>Libraries and SDKs. . . . .</b>	<b>144</b>
<b>9</b>	<b>Content Packages. . . . .</b>	<b>146</b>
<b>10</b>	<b>Advanced Features. . . . .</b>	<b>147</b>
10.1	AI Content as a Service. . . . .	147
	Service Custom Resource. . . . .	149
	Getting Started as a Service Provider. . . . .	151
	Metering. . . . .	153
	Offboarding. . . . .	153
	Shared Resource Group. . . . .	154
<b>11</b>	<b>Security. . . . .</b>	<b>157</b>
11.1	Security Features of Data, Data Flow, and Processes. . . . .	157
11.2	Encryption in Transit. . . . .	157

11.3	Authentication and Administration. . . . .	158
11.4	Docker Images. . . . .	158
11.5	AI Content Security. . . . .	158
11.6	Kubernetes Security. . . . .	159
11.7	Configuration Data and Secrets. . . . .	160
11.8	Output Encoding. . . . .	161
11.9	Multitenancy. . . . .	161
11.10	Auditing and Logging Information. . . . .	162
11.11	Data Protection and Privacy. . . . .	165
	Data Storage and Processing. . . . .	165
	Change Logging and Read-Access Logging. . . . .	165
	Consent. . . . .	166
	Deletion. . . . .	166
	Security and Customer Data Protection. . . . .	166
<b>12</b>	<b>Accessibility Features in SAP AI Core. . . . .</b>	<b>167</b>
<b>13</b>	<b>Monitoring and Troubleshooting. . . . .</b>	<b>168</b>
13.1	Troubleshooting. . . . .	168
	Repository. . . . .	169
	Configuration. . . . .	170
	Artifacts. . . . .	172
	Application. . . . .	174
	Execution. . . . .	179
	Docker. . . . .	181
	Deployment. . . . .	182
	Miscellaneous. . . . .	183
<b>14</b>	<b>Support Process. . . . .</b>	<b>186</b>
<b>15</b>	<b>Service Offboarding. . . . .</b>	<b>187</b>

# 1 What Is SAP AI Core?

Learn more about the SAP AI Core service on SAP Business Technology Platform (SAP BTP). Build a platform for your artificial intelligence solutions.



SAP AI Core is a service within the SAP Business Technology Platform. It's designed to manage the execution and operations of AI assets in a standardized, scalable, and hyperscaler-agnostic manner. It seamlessly integrates with SAP solutions, allowing any AI function to be easily implemented using open-source frameworks. SAP AI Core supports full lifecycle management of AI scenarios. Users can access generative AI capabilities and prompt lifecycle management through the generative AI hub.

SAP AI Core lets you make data-driven decisions confidently and efficiently, tailored to address business challenges. It handles large volumes of data and offers scalable machine learning capabilities to automate tasks like triaging customer feedback or tickets and performing classification tasks. SAP AI Core includes preconfigured SAP solutions and supports open-source machine learning frameworks. It integrates with Argo Workflow and KServe, and can be embedded into other applications.

SAP AI Core allows you to experiment with and utilize natural language prompts with a variety of generative AI models in the generative AI hub.



## → Tip

The English version of this guide is open for contributions and feedback using GitHub. This allows you to get in contact with responsible authors of SAP Help Portal pages and the development team to discuss documentation-related issues. To contribute to this guide, or to provide feedback, choose the corresponding option on SAP Help Portal:

- [▶ Feedback ▶ Create issue](#) : Provide feedback about a documentation page. This option opens an issue on GitHub.
- [▶ Feedback ▶ Edit page](#) : Contribute to a documentation page. This option opens a pull request on GitHub.

You need a GitHub account to use these options.

More information:

- [Contribution Guidelines](#)
- [Introduction Video](#) 
- [Introduction Blog Post](#) 

## Features

### Generative AI hub

Access generative AI models for prompt development and experimentation.  
Manage the lifecycle and evaluation of prompts and generative AI workflows.

<b>Execute pipelines</b>	Execute pipelines as a batch job, for example, to preprocess or train your models, or perform batch inference.
<b>Serve inference requests</b>	Deploy a trained machine learning model as a Web service to serve inference requests of trained models with high performance.
<b>Manage the AI scenario lifecycle</b>	Manage your ML artifacts and workflows, such as model training, metrics tracking, data, models, and model deployments via a uniform API lifecycle.
<b>Benefit from multitenancy support</b>	Use this service in tenant-aware applications. Implement multi-tenant services to segregate your AI assets and executions to isolate your tenants within SAP AI Core.
<b>Integrate your cloud infrastructure</b>	Register your Docker registry, synchronize your AI content from your git repository, and register your object store for training data and trained models. Productize your AI content and expose it as a service to consumers in the SAP BTP marketplace.

## Environment

This service is available in the following environments:

- Cloud Foundry
- Kyma
- Kubernetes

## Multitenancy

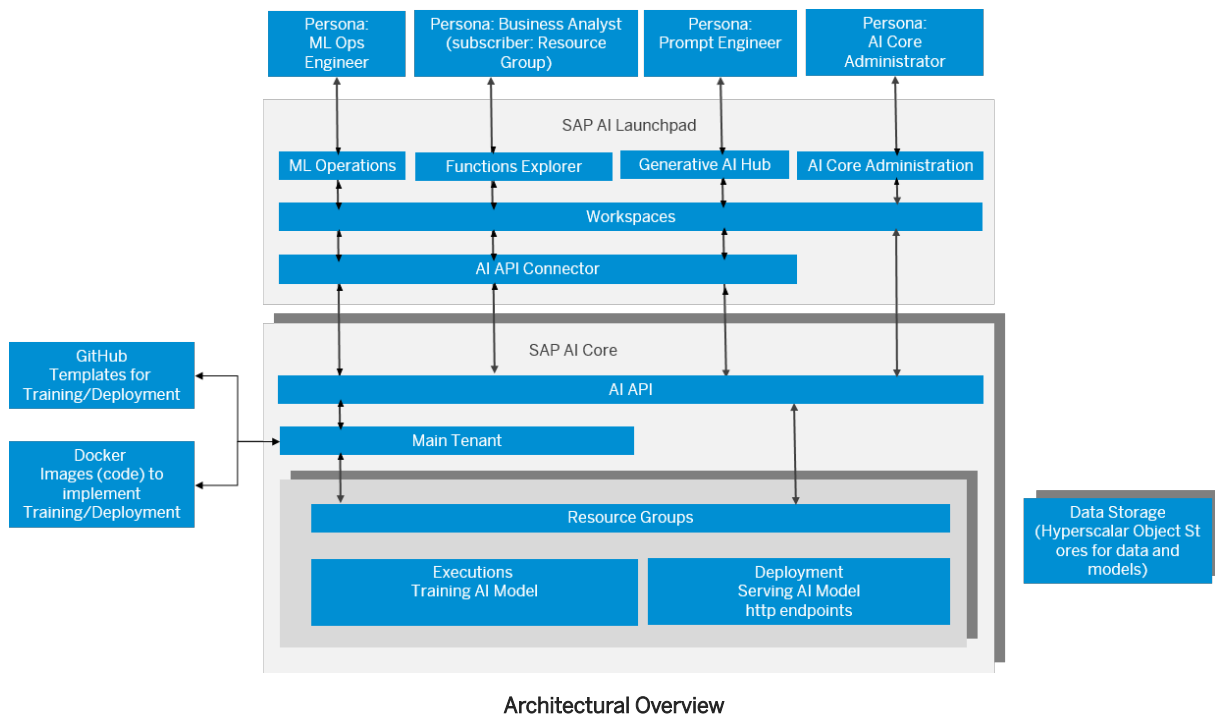
This service supports multitenancy. It can be used in tenant-aware applications. For more information, see [Multitenancy \[page 161\]](#).

## Architectural Overview

The SAP AI Core service works in conjunction with SAP AI Launchpad and the AI API. The key components are SAP AI Core, SAP AI Launchpad, and the AI API.

- **SAP AI Core** provides an engine that lets you run AI workflows and model serving workloads.
- **SAP AI Launchpad** manages a number of AI runtimes. It allows various user groups to access and manage their AI scenarios.
- **AI API** provides a standard way of managing the AI scenario lifecycle on different runtimes, regardless of whether they're provided on SAP technology (such as SAP S/4HANA) or on partner technology (such as Amazon Web Services). When the AI API is deployed on runtimes other than SAP AI Core, the runtimes have to provide a runtime adapter.

[Architectural Overview \[page 7\]](#) shows these three main components in an overview architecture diagram.



## Tools

SAP AI Core connects with various internal and external tools. You interact with different repositories, systems, and objects. Some of these objects come from SAP, while others must be provided by you. This setup enables enhanced control through authorizations, and supports continuous integration and continuous deployment (CI/CD). The following table lists the key repositories, systems, and objects:

What	Why
AI API	For managing your artifacts and workflows (such as training scripts, data, models, and model servers) across multiple runtimes
<p><b>Note</b></p> <p>The AI API can also be used to integrate other machine learning platforms, engines, or runtimes into the AI ecosystem.</p>	
Argo Workflows	A container native, workflow engine for Kubernetes.
Docker repo	For custom Docker images referenced in the templates
Git repo	For storing training and serving workflows and templates
Hyperscaler storage	For storage of input and output artifacts, such as training data and models (for example, SAP BTP Object Store Service)
KServing (K)	For optimized deployments of machine learning models. Deployment templates use KServe notation.
Kubernetes (K8s)	The K8s cluster orchestrates and scales the pods, which are used in AI pipelines. Resource group isolation is based on a K8s namespace.

What	Why
SAP AI Launchpad	SAP AI Launchpad is a multitenant software as a service (SaaS) application on SAP Business Technology Platform (SAP BTP). Customers and partners can use SAP AI Launchpad to manage AI use cases (scenarios) across multiple instances of AI runtimes (such as SAP AI Core). SAP AI Launchpad also provides generative AI capabilities via the Generative AI Hub.

## Related Information

[Generative AI Hub](#)

[AI API Overview \[page 55\]](#)

[SAP AI Launchpad](#)

## 1.1 Metering and Pricing for SAP AI Core

SAP AI Core provides a scalable infrastructure for AI model management, with usage-based pricing that lets you pay only for the resources you use.

The service offers various plans, including a free, standard, and extended plan. If you want to access all capabilities, we recommend the extended plan.

We offer the following model options:

- Generative AI with SAP-hosted, SAP-managed, and remote foundation models, and pipelines such as grounding
- Predictive AI with your custom AI models

You have the flexibility to select the model options that best suit your needs.

Custom AI models incur compute, storage, and baseline costs. The system automatically applies baseline charges every hour, up to a maximum of 730 hours.

If you use foundation models and grounding, the platform waives compute, storage, and baseline costs. Instead, charges accrue based on the number of tokens used by the foundation models.

## More Information

For more information about the potential costs associated with using foundation models and grounding in SAP AI Core, including metering information, example pricing calculations, and typical consumption patterns, see [Metering and Pricing for Generative AI \[page 9\]](#).

For more information about the potential costs associated with using custom AI models in SAP AI Core, including metering information, example pricing calculations, and typical consumption patterns, see [Metering and Pricing for Predictive AI \[page 13\]](#).

## Consumption Information

To see detailed consumption information in SAP BTP cockpit, open your global account and click [Usage](#) from the navigation panel. Filter by AI Core and click Export. The report contains generative AI hub model consumption under [Applications](#) and resource group level consumption under [instance](#).

For more information, see [Monitoring Usage and Consumption Costs in Your Global Account in SAP BTP](#).

### 1.1.1 Metering and Pricing for Generative AI

The use of generative AI in SAP AI Core is measured using tokens, which are converted into capacity units based on input and output volume. Billing is calculated according to these capacity units and varies by model type and usage pattern.

#### Note

The generative AI hub is available only as part of the extended service plan.

### Metering for Generative AI

Use of large language models (LLMs) is metered in tokens. Tokens are the fundamental building blocks of text that a language model processes. They represent single characters, parts of words, entire words, or even punctuation marks. They're the unit into which the LLM breaks down text input and output.

It's important to distinguish between input and output tokens. Input tokens refer to the text provided in the prompt, while output tokens represent the response generated by the LLM. The ratio between input and output tokens varies depending on the prompt and the use case. For example, in summarization tasks, the ratio of input to output tokens could be 3:1 or even 4:1. In contrast, tasks like generating explanations or drafting longer pieces of text could have an inverse ratio, with output tokens significantly outnumbering input tokens.

The number of tokens required depends on the content and complexity of your prompt and the specific task assigned to the LLM. Token consumption varies with the capabilities of the LLM being used, with more advanced models typically requiring and processing a higher number of tokens.

Consumption of generative AI in SAP AI Core is measured in GenAI tokens. GenAI tokens are a virtual unit used to represent usage across different models. The actual number of processed tokens (input and output) that corresponds to one GenAI token varies by model.

Generative AI resource requirements vary based on the use case. Available resources are shown in the table:

Resource Type	SAP AI Core Resources	Unit of Measure (UoM)
Provisioning of foundation models	Access to (LLMs) and other generative AI capabilities	Tokens

Resource Type	SAP AI Core Resources	Unit of Measure (UoM)
Baseline	Grounding Service	Gigabyte day
Inference Observability	Storing inference records	DataVolume

For more information on inference observability, including conversion rates between inference records and DataVolume units, see [3720903](#).

## Pricing for Generative AI

The billing metric for SAP BTP is referred to as “capacity units” (CUs). Capacity units are calculated using conversion factors applied to GenAI tokens.

In general, output tokens tend to be slightly more expensive than input tokens.

For more information about the supported models, including the conversion rates between model tokens and GenAI tokens, see SAP Note [3437766](#).

Here's an example estimation, using fictitious values:

### ❁ Example

This example contains:

- A retrieval-augmented generation (RAG) system handling 25,000 requests in a single month results in the following costs:
- Language model use: The total of 25,000 requests, each with 3,500 input tokens and 300 output tokens.
- Grounding service: The RAG pipeline retrieves and processes context using the following orchestration modules:
  - Storage in Grounding (for storing relevant documents)
  - Retrieval in Grounding (for fetching context during inference)
  - Content Filtering
  - Data Masking (for privacy and compliance)
- Inference Observability: The 25,000 inference requests and responses are stored with associated metadata, labels and feedback for later analysis.

The following orchestration services are essential to support the grounding step and introduce additional capacity usage beyond the core LLM calls.

- The Grounding storage module stores contextual documents in the vector engine.
- The Grounding retrieval module retrieves context documents during inference.
- Content Filtering via Meta Llama Guard helps to ensure content safety and relevance.
- Data Masking, handled by SAP Data Privacy Integration, applies masking at the API level.

Each module uses its own metric (storage size, text blocks, API calls) multiplied by monthly usage, then converted to capacity units to estimate resource consumption and cost.

**Calculate the calls incurred through LLM calls:**

This example uses a model with an input GenAI token conversion rate of 0.00112 per 1,000 input tokens. The output GenAI token conversion rate is 0.00320 per 1,000 output tokens.

1. Calculate Total GenAI tokens

Formula: Total GenAI tokens = (x/1000 × conversion rate per 1,000 input tokens) + (y/1000 × conversion rate per 1000 output tokens)

In this formula, “x” is the number of input tokens and “y” is the number of output tokens. The conversion rates per 1,000 input and output tokens are given constants.

This example includes 3,500 input tokens and 300 output tokens. The conversion rate is 0.00112 per 1,000 input tokens and 0.00320 per 1,000 output tokens.

Values: Total GenAI Tokens = (3500/1000 × 0.00112) + (300/1000 × 0.00320) = 0.00488

2. Calculate capacity units

Formula: capacity units = Total GenAI Tokens × 1.90385

In this formula, “Total GenAI Tokens” is the value calculated in the previous step and 1.90385 the capacity unit value, given as a constant.

Values: capacity units = 0.00488 × 1.90385 = 0.00929

3. Calculate Total Usage for handling 25,000 requests

Formula: Total capacity units = Requests × capacity units per request

In this formula, “Requests” is the number of requests to be handled (in this case, 25,000) and “capacity units per request” is the value calculated in the previous step.

Values: Total capacity units = 25,000 × 0.00929 = \$232.25

**Calculate Grounding Module Costs:**

The RAG architecture needs various supporting services to handle contextual data for the language model. These services use resources measured by specific metrics, which are converted into capacity units based on usage volume and internal conversion factors

Each module uses its own metric (storage size, text blocks, API calls) multiplied by monthly usage. The values are converted to capacity units to quantify resource consumption and estimate cost.

Module	Service	Metric	Metric Value	Occurrences / Month	Capacity Units / Month
Grounding	Storage	Storage in GB / Day	1	30	6.2 CU
Grounding	Retrieval	Text Blocks	1	25.000	30 CU
Content Filter	Meta Llama Guard	Text Blocks	5	25.000	84 CU
Data Masking	SAP Data Privacy Integration	API Calls	1	25.000	66.1 CU
Total					186.3 CU

Total consumption = 232.25 + 186.3 = 418.55 CU

### Calculate Inference Observability Costs:

This example assumes that an inference record consumes 0.002048 MB per day. We assume all 25,000 records are created at day one and are kept for a month.

1. Calculate total storage consumption  
Formula: Total storage consumption per day = Number of inference records \* 0.002048 MB  
Values: Total storage consumption per day = 25,000 \* 0.002048 MB = 51.2 MB
2. Calculate data volume units per day  
Formula: data volume units per day = total storage consumption per day (in MB) \* 0.00089  
Values: data volume units per day = 51.2 \* 0.00089 = 0.045568
3. Calculate capacity units per month  
Formula: capacity units per month = data volume units per day \* 30  
Values: capacity units per month = 0.045568 \* 30 = 1.36704 CU

## Images

For image use with Mistral Small Instruct, images are converted into tokens in batches of 14\*14 pixels. The formula for the number of image tokens is:  $(\text{ResolutionX}/14) * (\text{ResolutionY}/14)$ . Images with a resolution higher than 1540\*1540 are downscaled, while maintaining their aspect ratio.

### ❁ Example

This example uses fictitious values.

Given an image of resolution 720\*512, the number of GenAI tokens consumed through the image is:

$$\text{GenAI Image tokens} = (720/14) * (512/14) = 1,880$$

The calculation for image tokens also includes additional sequence start and end tokens, as well as tokens for image breaks.

## Typical Consumption Patterns

The following table displays API request counts for productive IT use cases based on the consumption patterns "small", "medium", and "large". It also specifies the average distribution of input and output tokens, which can help you estimate potential costs for consuming foundation models.

Use Case	Input Tokens	Output Tokens	Tokens Occurrences / Request per Month: Small	Tokens Occurrences / Request per Month: Medium	Tokens Occurrences / Request per Month: Large
RAG Chat	3500	300	1000	25k	300k
Basic Chat	500	100	1000	30k	300k

Use Case	Input Tokens	Output Tokens	Tokens Occurrences / Request per Month: Small	Tokens Occurrences / Request per Month: Medium	Tokens Occurrences / Request per Month: Large
Summarization	5000	300	1000	25k	300k
Classification	3800	10	1000	5k	50k
Generation	500	3500	1000	2k	50k

The generative AI hub provides additional modules through the Orchestration API, including:

- Grounding
- Content Filtering
- Data Masking

Charges associated with the use of other SAP AI Core components can also apply. For more information, see [Metering and Pricing for Predictive AI \[page 13\]](#).

#### → Recommendation

For an estimate of projected costs, use the [SAP AI Core Cost Calculator](#).

## 1.1.2 Metering and Pricing for Predictive AI

### Metering for Predictive AI

Custom AI workloads involve diverse resource types, each with varying performance and efficiency levels. Measuring key resources like compute power, storage, and baseline costs is crucial. Non-billable units of measure (UoM) are used to represent the consumption of these resources.

For example, compute-heavy AI models on high-performance instances consume more resources per hour than smaller, less demanding workloads. This results in proportionally higher costs. Pricing reflects true resource consumption.

Custom AI development requirements vary significantly based on the use case. A range of compute infrastructure resources is available, differing in CPU cores, memory (GB), and GPU presence for GPU-powered workloads like model training. Available resources are shown in the table.

#### ⓘ Note

Infrastructure specifications depend on the chosen hyperscaler. For more information, see [Choose an Instance](#).

Resource Type		SAP AI Core Resources	Unit of Measure (UoM)
Compute	Instance Types	For information about available instance types, see SAP Note <a href="#">3660109</a> .	Node hour
	Resource Plans	Starter Instance (min. 2.5 GB, 1 CPU core)	
		Basic Instance (min. 10 GB, 3 CPU cores)	
		Basic.8x Instance (min. 115 GB, 31 CPU cores)	
		Infer-S Instance (min. 10 GB, 3 CPU cores, standard GPU)	
		Infer-M Instance (min. 25 GB, 7 CPU cores, standard GPU)	
		Infer-L Instance (min. 55 GB, 15 CPU cores, standard GPU)	
	Train-L Instance (min. 47 GB, 5 CPU cores, advanced GPU)		
Storage		Standard SSD Volume	Gigabyte hour
Baseline		Fixed cluster resources (instances, storage, SAP HANA, Kubernetes, DevOps, support, and so on)	Tenant hour

Custom AI models can incur compute, storage, and baseline costs. You have the flexibility to select the components that best suit your needs. If you use compute or storage resources, the system automatically applies baseline charges every hour, up to a maximum of 730 hours.

#### **Note**

You can process data in parallel by using multiple nodes. Costs are calculated in node hours. For example, processing data for 1 hour on 2 nodes equals 2 node hours. However, it incurs only 1 hour of baseline costs.

## Pricing for Predictive AI

The billing metric for SAP BTP is referred to as “capacity units”. Capacity units are calculated using conversion factors applied to compute, storage, and baseline intermediary UoM values. This allows you to calculate a specific price for the infrastructure used based on your compute and storage requirements.

Here's an example calculation, using fictitious values:

### 🔗 Example

Training a small model on a GPU node and serving it on a CPU node later in the same month results in the following costs:

- Baseline: Total node hours of compute instances across the full month
- Storage: Temporary high-speed storage (for example, SSD) used during training and inference to stream or copy data
- Compute: Compute consumption based on node hours and chosen compute instances across the full month

Resource Type	Unit of Measure	Capacity Unit (CU)	
		Value	Fictional Usage
Compute	node hour	0.3900	300 node hours in Basic Instance
Compute	node hour	1.552	100 node hours in Infer-M Instance
Storage	gigabyte hour	0.0003	10.000 gigabyte hours
Baseline	tenant hour	1.2241	400 baseline tenant hours

Infrastructure specifications are combined minimum values; actual numbers vary by deployed hyperscaler and may be higher.

The CUs can be converted to monetary values in the BTP Discovery Center Estimator. For more information, see [BTP Discovery Center Estimator](#).

### → Recommendation

For an estimate of projected costs, use the [SAP AI Core Cost Calculator](#).

## Typical Consumption Patterns

Depending on your infrastructure and your consumption pattern (S, M or L), you need on average the following node hours for compute, storage, and baseline hours.

T Shirt Size	Starter	Basic	Basic8x	Infer.S	Infer.M	Infer.L	Train.L	Storage	Baseline
S	50	30	0	300	0	0	50	0	430

<b>T Shirt</b>									
<b>Size</b>	<b>Starter</b>	<b>Basic</b>	<b>Basic8x</b>	<b>Infer.S</b>	<b>Infer.M</b>	<b>Infer.L</b>	<b>Train.L</b>	<b>Storage</b>	<b>Baseline</b>
M	0	900	0	160	1880	1440	120	0	730
L	0	0	0	800	10000	0	2300	0	730

## 2 What's New for SAP AI Core

Tech	Envi-								Mod			
nical	ron-								ular			
Com	men								Busi			
po-									ness			
nent	nt	Title	Description	Ac-	Life-				Proc			
ent				tion	cy-	Type	of	ness	ess	Product	Lat-	Avail
					cle		Busi				est	able
							ness				Revi-	as of
											sion	
SAP AI Core	Cloud Foundry	Generative AI hub	New models are supported, including Gemini 3.1 Flash Lite.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	General Availability	Announcement	Technology	Not applicable	SAP Business Technology Platform	2026-05-12	2026-05-12	
SAP AI Core	Cloud Foundry	Generative AI hub	For model consumption through foundation models, you can reduce costs and by sending requests in batches.  For more information, see <a href="#">Batch Consumption</a> .	Info only	General Availability	Announcement	Technology	Not applicable	SAP Business Technology Platform	2026-05-08	2026-05-08	
SAP AI Core	Cloud Foundry	Generative AI hub	<code>metaConfigId</code> has been added as a filter for pipeline queries in the pipelines API for the grounding module of orchestration.  For more information, see <a href="#">Data Pipelines</a> .	Info only	General Availability	Announcement	Technology	Not applicable	SAP Business Technology Platform	2026-05-08	2026-05-08	
SAP AI Core	Cloud Foundry	Generative AI hub	The templating module of orchestration has been updated to include prompt caching for use with OpenAI and Gemini models only.  Prompt caching improves performance by reusing common prompt sections, enabling faster and more consistent responses across requests.  For more information, see <a href="#">Prompt Caching</a> .	Info only	General Availability	Announcement	Technology	Not applicable	SAP Business Technology Platform	2026-05-08	2026-05-08	

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- noun- ce- men- t Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	The templating module of or- chestration has been updated to include response formatting.  You can define schemas for the LLM to follow to guide the struc- ture of your outputs.  For more information, see <a href="#">Re- sponse Formatting</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-05 -08	202 6-05 -08
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can now track the prog- ress of long-running prompt op- timizations in real-time. When viewing execution status, prog- ress details are automatically displayed in a normalized for- mat.  You can also use additional met- rics such as ROUGE, BLEU, COMET and metrics designed to evaluate tool calling.  For more information, see <a href="#">Prompt Optimization</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-05 -08	202 6-05 -08
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding Claude Opus 4.7, GPT Realtime and Mistral Small. Ad- ditionally, Gemini-2.5 Flash sup- port image generation.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-05 -08	202 6-05 -08

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Inference observability allows you to store and retrieve payloads from foundation model inferences for analytical purposes. You can add custom labels for filtering, attach feedback after inference completion, and store data in a registered S3 object store or opt for metadata-only storage. Use inference headers to explicitly control which requests are recorded, and leverage the REST APIs to manage labels, feedback, and retrieve records for analysis.  For more information, see <a href="#">Inference Observability</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-04 -27	202 6-04 -27
SAP AI Core	Clou- d Foun- dry	Generative AI hub	The Google Drive object store has been added to orchestration workflows.  For more information, see <a href="#">Create a Generic Secret for Google Drive</a> and <a href="#">Create a Pipeline with Google Drive</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-04 -27	202 6-04 -27
SAP AI Core	Clou- d Foun- dry	Predictive AI Execu- tions	You can now track the progress of long-running executions in real-time. When viewing execution status, progress details are automatically displayed in a normalized format.  For more information, see <a href="#">Start Training</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-04 -27	202 6-04 -27

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Administra- tion	mTLS certificate secrets are supported, enabling your AI workloads to securely authenticate to external services using mutual TLS (mTLS) without manual certificate management.  For more information, see <a href="#">Manage mTLS Certificate Secrets [page 130]</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-04 -27	202 6-04 -27
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, including GPT 5.4, GPT 5.4-nano and GPT 5.3-Codex.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-04 -27	202 6-04 -27
SAP AI Core	Clou- d Foun- dry	Generative AI hub	For security reasons, the system prompt for SAP-ABAP-1 is now predefined. The prompt is added automatically to all requests.  If you include a system prompt in a request, the request fails and an error is returned.  In most scenarios, a user prompt is sufficient. If you need additional instructions, include them at the beginning or end of the user prompt.	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-03 -01	202 6-03 -01

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	The <code>masking_providers</code> property in the Orchestration V2 Data Masking module configuration is deprecated and has been replaced by the <code>providers</code> property. If you are currently using <code>masking_providers</code> in your masking module configurations, update your specifications to use the <code>providers</code> property instead. The deprecated <code>masking_providers</code> property will be removed on September 15, 2026.  For more information, see <a href="#">Enhancing Model Consumption with Data Masking</a> .	Info only	Dep- re- cate- d	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-03 -16	202 6-03 -16
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, including Amazon Nova Lite 2.0.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-03 -16	202 6-03 -16
SAP AI Core	Clou- d Foun- dry	Generative AI hub	SAP RPT 1 supports <code>top_k</code> as an input parameter, and outputs now include confidence intervals.  For more information, see <a href="#">Example Payloads for Inferencing: sap-rpt-1</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-03 -01	202 6-03 -01
SAP AI Core	Clou- d Foun- dry	Generative AI hub	PDF documents are supported in the Data Masking module in Orchestration, and Service Now has been added as a grounding repository.  For more information, see <a href="#">Enhancing Model Consumption with Data Masking and Generic Secrets for Grounding</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-03 -01	202 6-03 -01




Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- nounce- ment Type	Tech nol- ogy	Mod- ular Busi- ness Proc- ess	Line of Busi- ness Product	Lat- est	Avail- able
										Revi- sion	as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, including Anthropic Claude Opus 4.6.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- nounce- ment	Tech nol- ogy		SAP Business Technology Platform	202 6-03 -01	202 6-03 -01
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can now manage metadata for your documents, collections and chunks created using the Vector API, to enable advanced filtering and organization of your content.  For more information, see <a href="#">Meta-data</a> .	Info only	Gen- eral Avail- abil- ity	An- nounce- ment	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-02 -16	202 6-02 -16
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can merge and rank search results from multiple data repositories using post-processing capabilities of the Retrieval API.  For more information, see <a href="#">Retrieval Search</a> .	Info only	Gen- eral Avail- abil- ity	An- nounce- ment	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-02 -16	202 6-02 -16
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Vector search supports advanced filtering capabilities including metadata filtering.  For more information, see <a href="#">Vector Search</a> .	Info only	Gen- eral Avail- abil- ity	An- nounce- ment	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-02 -16	202 6-02 -16
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Custom metrics are supported in prompt optimizations, enabling you to define and optimize prompts based on your specific evaluation criteria. Only LLM-as-a-judge metrics with numerical or Boolean output types can be used in optimization tasks.  For more information, see <a href="#">Create a Custom Metric</a> and <a href="#">Create a Configuration for a Prompt Optimization</a> .	Info only	Gen- eral Avail- abil- ity	An- nounce- ment	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-02 -16	202 6-02 -16

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Tech of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Custom metric creation has been enhanced.  For more information, see <a href="#">Create a Custom Metric</a> .	Info only	Dep- re- cate d	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-02 -16	202 6-02 -16
SAP AI Core	Clou- d Foun- dry	Generative AI hub	The Prompt Registry now ena- bles you to create and manage orchestration configurations de- claratively, allowing you to ver- sion and track complex AI work- flows alongside your prompts for better governance and reproduc- ibility.  For more information, see <a href="#">Create an Orchestration Config (Im- perative)</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-02 -16	202 6-02 -16
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding Open AI GPT 5.2..  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-02 -16	202 6-02 -16
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can specify the <code>promptTemplateScope</code> ac- cess level in evaluation and prompt optimization configura- tions.  For more information, see <a href="#">Create an Evaluation</a> and <a href="#">Create a Configuration for a Prompt Opti- mization</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-01- 30	202 6-01- 30
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can provide separate test and train datasets for use in prompt optimizations.  For more information, see <a href="#">Create a Configuration for a Prompt Optimization</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-01- 19	202 6-01- 19

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	When using the Vector API, you can assign a UUID during collection creation and bulk delete documents by ID.  For more information, see <a href="#">Create a Collection</a> and <a href="#">Delete Documents</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-01- 19	202 6-01- 19
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can now use the Claude Opus 4.5 model with Amazon Bedrock. The model is available with the extended service plan.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t		Not ap- pli- ca- ble	SAP Business Technology Platform	202 6-01- 19	202 6-01- 19
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can now use the SAP-ABAP-1 model to better understand ABAP code. SAP-ABAP-1 is a foundation model built by SAP and is fine-tuned on a large amount of ABAP code.  For more information, see <a href="#">SAP-ABAP-1</a> and <a href="#">Example Payloads for Inferencing: SAP-ABAP-1</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -12-2 2	2025 -12-2 2

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	<p>SAP-RPT-1 launched.</p> <p>SAP-RPT-1 is a relational pre-trained transformer for use on relational and structured data. It is developed and maintained by SAP.</p> <p>Relational Foundation Models (RFMs) are large-scale machine learning models designed to understand, process, and do predictions on tabular and relational data.</p> <p>RPT-1 solves predictive tasks such as classification and regression out-of-the-box without requiring any training or fine-tuning via in-context learning. Due to its table-native architecture, prediction quality on enterprise data is typically very high, ahead of state-of-the-art narrow AI models and LLMs employed for such tasks.</p> <p>For more information, see <a href="#">SAP-RPT-1</a> and <a href="#">Example Payloads for Inferencing: sap-rpt-1</a>.</p>	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- cable	SAP Business Technology Platform	2025- 12-08	2025- 12-08

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	<p>Evaluations added to Optimizations</p> <p>Evaluations provides tools for benchmarking large language models and prompts via orchestration configurations.</p> <p>Evaluations can be used for the following use cases:</p> <ul style="list-style-type: none"> <li>Evaluating prompt templates and models as orchestration configurations to determine the most effective combination for a use case.</li> <li>Use industry standard pre-defined metrics for model and prompt comparison with your use case specific dataset.</li> <li>Use your own custom defined metrics for your prompt and model evaluation.</li> </ul> <p>For more information, see <a href="#">Evaluations</a>.</p>	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -12-0 8	2025 -12-0 8
SAP AI Core	Clou- d Foun- dry	Generative AI hub	<p>Metrics added to Optimizations</p> <p>You can choose from a selection of system defined metrics, or define your own custom llm-as-a-judge metrics, including rating criteria. You can incorporate system defined and custom metrics into your evaluation workflows, and manage the lifecycle of your custom metrics.</p> <p>For more information, see <a href="#">Metrics</a>.</p>	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -12-0 8	2025 -12-0 8

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Tech of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding Anthropic Claude 4.5 Haiku.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -12-08	2025 -12-08
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding Gemini 2.5 Flash Lite.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -12-01	2025 -12-01
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Orchestration V2 includes sup- port for a unified embeddings endpoint that works consistently across different LLM providers. This endpoint can be used inde- pendently or combined with data anonymization capabilities.  For more information, see <a href="#">Em- beddings</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -11-07	2025 -11-07
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can define multiple model configurations in your orches- tration workflow, enabling auto- matic fallbacks if processing fails (for example, due to unsup- ported models in a region).  For more information, see <a href="#">Or- chestration with Fallbacks</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -11-07	2025 -11-07
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding Perplexity Sonar and So- nar Pro.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -10-00	2025 -10-00

Tech nical Com- po- nent	Envi- ron- men t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Tech of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est	Avail- able
										Revi- sion	as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can optimize prompts to by evaluating them against a specified metric. The optimized prompt is saved, improving model performance and output quality with prompt reuse.  For more information, see <a href="#">Prompt Optimization</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -10-2 0	2025 -10-2 0
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can manage the lifecycle of your orchestration configs, in- cluding creating, saving, version- ing, reuse. CRUD functionality, and import and export of or- chestration configs are also sup- ported.  For more information, see <a href="#">Or- chestration Config Management</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -10-2 0	2025 -10-2 0
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can use the quota manage- ment API endpoints to check and update your rate limits for models in the generative AI hub.  For more information, see <a href="#">Rate Limit Management</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -10-0 6	2025 -10-0 6
SAP AI Core	Clou- d Foun- dry	Generative AI hub	SAP Document Management is supported as a repository type for the orchestration grounding module.  For more information, see <a href="#">Grounding Generic Secrets for SAP Document Management Service</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -09- 01	2025 -09- 01
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Orchestration model configura- tions have timeout and max tries parameters.  For more information, see <a href="#">Tem- plating</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -09- 01	2025 -09- 01

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding GPT 5, GPT 5 Mini, and GPT 5 Nano.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -08-1 8	2025 -08-1 8
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding Gemini Embeddings.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -08- 06	2025 -08- 06
SAP AI Core	Clou- d Foun- dry	Generative AI hub	This initial orchestration end- point is deprecated and is scheduled for decommissioning on October 31, 2026. Following that date, the endpoint will no longer be available. We recom- mend that you create new or- chestration workflows using ver- sion 2, and that you migrate existing workflows from version 1 to version 2. To use version 2, you'll need to update your endpoint from <code>/completion</code> to <code>/v2/completion</code> and modify your existing payloads. For more information, see SAP Note <a href="#">Orchestration Workflow V2</a> and <a href="#">3634540</a> .	Info only	Dep- re- cate- d	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -07- 21	2025 -07- 21
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding Nova Premier from AWS Bedrock.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -07- 21	2025 -07- 21



Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Prompt shields are available as part of input filtering with Azure Content Safety. Prompt shields detect and mitigate prompt attacks, such as prompts designed to bypass safety mechanisms or override previous instruction.  For more information, see <a href="#">Harm categories in Azure AI Content Safety</a> and <a href="#">Input Filtering</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -06- 23	2025 -06- 23
SAP AI Core	Clou- d Foun- dry	Generative AI hub	GPT-4 versions 0613 and turbo-2024-04-09 and gpt-4-32k version 0613 have been deprecated.  It's recommended that users of these models migrate to GPT-4.1 as a replacement.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Dep- re- cate- d	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -06- 02	2025 -06- 02
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Gemini-1.5-pro and gemini-1.5-flash have been deprecated.  It's recommended that users of these models migrate to gemini-2.0-flash, or gemini-2.0-flash-lite as replacements.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Dep- re- cate- d	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -06- 02	2025 -06- 02
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Mistralai--mixtral08x7b-v01 has been deprecated.  It's recommended that users of this model migrate to mistralai--mistral-small-instruct as a replacement.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Dep- re- cate- d	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -06- 02	2025 -06- 02

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Tech of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding OpenAI GPT, o3, o4-mini, 4.1, 4.1-mini, 4.1-nano and Mis- tral small instruct.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -05-1 9	2025 -05-1 9
SAP AI Core	Clou- d Foun- dry	Generative AI hub	An additional module has been added to orchestration.  The translation module trans- lates text, and can be configured for input and output text.  The input translation module helps improve answer quality when the configured model per- forms better when input is pro- vided in a specific language, for example English.  For more information, see <a href="#">Input Translation</a> and <a href="#">Output Transla- tion</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -04- 24	2025 -04- 24
SAP AI Core	Clou- d Foun- dry	SAP BTP Cockpit	A detailed consumption report is available in SAP BTP for Generative AI hub as the part of SAP AI Core service. You can see the consumption specific to an LLM model and Orchestra- tion service, and consumption for each resource group.  For more information, see <a href="#">Me- tering and Pricing for Generative AI [page 9]</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -04-1 4	2025 -04-1 4




Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est	Avail- able
										Revi- sion	as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	New models are supported, in- cluding Anthropic Claude 3.7 sonnet, GCP Vertex AI Gem- ini 2.0-flash and flash-lite, NVI- DIA Llama 3.2 nv embedqa 1b, and OpenAI GPT 4o version 2024-11-20.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -04- 4	2025 -04- 4
SAP AI Core	Clou- d Foun- dry	Generative AI hub	We now support the Anthropic models Claude Opus 4 and Claude Sonnet 4.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -06- 23	2025 -06- 23
SAP AI Core	Clou- d Foun- dry	Generative AI hub	More endpoints have been added to the Pipeline API, so you can retrieve information about the documents within a pipeline, including processing status.  For more information, see <a href="#">Data Pipelines</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -02- 03	2025 -02- 03
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Selected models from Aleph AI- pha are supported.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -02- 03	2025 -02- 03
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models and model versions from Open AI are supported.  For more information, see <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -03- 03	2025 -03- 03






Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	You can now configure your orchestration workflow so that Llama Guard 3 filters input and output content. Llama Guard 3 provides several content categories for filtering content.  Azure Content Safety is still supported.  For more information, see <a href="#">Input Filtering and Output Filtering</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2025 -01-1 9	2025 -01-1 9
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Prompt registry  Prompt registry integrates prompt templates into SAP AI Core, making them discoverable across your applications and orchestration. It reduces the complexity of dealing with prompt templates and leveraging integration capabilities.  For more information, see <a href="#">Prompt Registry</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -12-0 2	2024 -12-0 2
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Retrieval and vector APIs added.  The vector API can be used for managing collections and documents in the Vector database.  The retrieval API searches data repositories and returns the relevant chunks for the user query.  For more information, see <a href="#">Preparing Data Using the Vector API and Retrieval API</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -12-0 2	2024 -12-0 2

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models and model versions from Anthropic via AWS Bedrock are supported.  For more information, see <a href="#">Foundation Models</a> and SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -12-0 2	2024 -12-0 2
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models and model versions from Open AI are supported.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -12-0 2	2024 -12-0 2
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected model ver- sions from Vertex AI are sup- ported.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -12-0 2	2024 -12-0 2
SAP AI Core	Clou- d Foun- dry	Generic Se- crets	Generic secrets are now avail- able at the tenant-wide level.  For more information, see <a href="#">Cre- ate a Generic Secret [page 118]</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -12-0 2	2024 -12-0 2
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Pipeline API added.  This pipeline segments data into chunks and generates em- beddings, which are multidimen- sional representations of textual information. The embeddings are stored in a vector database.  For more information, see <a href="#">Cre- ate a Document Grounding Pipe- line Using the Pipelines API</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -11-0 4	2024 -11-0 4

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	A new module has been added to the orchestration system. The grounding module enhances AI processes by integrating external data that is contextually relevant, domain-specific, or real-time. This additional data complements the natural language processing abilities of pre-trained models, which are typically based on general information.  For more information, see <a href="#">Grounding</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -11-0 4	2024 -11-0 4
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models from IBM are supported.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -10-0 7	2024 -10-0 7
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models from Mistral AI are supported.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -10-0 7	2024 -10-0 7
SAP AI Core	Clou- d Foun- dry	Generative AI hub	For selected models, response streaming is supported in the generative AI hub orchestration service.  For more information, see <a href="#">Streaming</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -10-0 7	2024 -10-0 7

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	When creating a deployment for orchestration, you can restrict the choice of models using an allow or disallow list. You can use this to implement internal standards, for example where only certain LLMs are approved for use.  For more information, see <a href="#">Create a Deployment for Orchestration</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -09-1 6	2024 -09-1 6
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional module added to orchestration.  The data masking module anonymizes or pseudonymizes personally identifiable information from input.  For more information, see <a href="#">Orchestration Workflow V1 (Deprecated)</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -09-1 6	2024 -09-1 6
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models from Meta are supported.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -09-1 6	2024 -09-1 6
SAP AI Core	Clou- d Foun- dry	Generative AI hub	<code>tiiaue--falcon-40b-instruct</code> is deprecated and is unavailable from 2024-09-01. Users should choose another model.  For more information, see SAP Note <a href="#">3437766</a> .	Info only	Dep- re- cate- d	Cha- nged	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -09- 01	2024 -09- 01

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Tech of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models from OpenAI are supported via Azure.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -09- 02	2024 -09- 02
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models from Anthropic are supported via AWS Bedrock.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble		2024 -05- 20	2024 -05- 20
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Orchestration combines content generation with a set of functions that are often required in business AI scenarios.  Functions include templating, which lets you compose a prompt with placeholders that are filled during inference, and content filtering, which lets you restrict the type of content that is passed to and received from a generative AI model.	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -08- 05	2024 -08- 05
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models from GCP Vertex AI are supported.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -07- 08	2024 -07- 08
SAP AI Core	Clou- d Foun- dry	Resource limits per tenant	The maximum number of applications, git repository secrets, docker registry secrets, workflow templates and deployment templates is limited at tenant level to 50 per resource. If you reach this limit, you will receive an error message. You can free up space by deleting resources.	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -06- 24	2024 -06- 24

Tech nical Com- po- nent	Envi- ron- men t	Title	Description	Ac- tion	Life- cy- cle	An- noun- ce- men t	Tech nol- ogy	Mod- ular Busi- ness Proc- ess	Line of Busi- ness Product	SAP Business Technology Platform	Lat- est	Avail- able
											Revi- sion	as of
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Additional selected models from OpenAI are supported via Azure.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -06- 24	2024 -06- 24	
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Selected models from Amazon and Anthropic are supported via AWS Bedrock.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -06- 03	2024 -06- 03	
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Selected models from Meta are supported.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -06- 03	2024 -06- 03	
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Selected models from Mistral AI are supported.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -05- 20	2024 -05- 20	
SAP AI Core	Clou- d Foun- dry	Deploy- ments	When deployments are submitted, configurations are checked for errors, synchronously.  For more information, see <a href="#">Deploy Models</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -04- 08	2024 -05- 20	
SAP AI Core	Clou- d Foun- dry	Generative AI hub	Selected models from GCP Vertex AI are supported.  For more information, see SAP Note <a href="#">3437766</a>  .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -04- 08	2024 -04- 08	

Tech nical Com- po- nent	Envi- ron- men t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou d Foun dry	Resource groups limit per tenant	The maximum number of re- source groups is limited at ten- ant level to 50. If you reach this limit, you receive an error mes- sage. To free up space, delete some resource groups. Alterna- tively, raise a ticket to increase your quota.  For more information, see <a href="#">Delete a Resource Group [page 100]</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -03- 08	2024 -03- 08
SAP AI Core	Clou d Foun dry	Authentica- tion	Authentication through X.509 certificates is supported.  For more information, see <a href="#">Cre- ate a Service Key [page 77]</a> and <a href="#">Use a Service Key [page 80]</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -02-1 8	2024 -02-1 8
SAP AI Core	Clou d Foun dry	generative AI hub SDK	The generative AI hub SDK is now available.  For more information, see <a href="#">gener- ative AI hub SDK</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2024 -02-1 8	2024 -02-1 8
SAP AI Core	Clou d Foun dry	Generative AI hub	The generative AI hub in- cludes prompt experimentation, prompt management, and ad- ministrative tools. Prompt ex- perimentation includes creating and running natural language prompts with a choice of large language models and paramet- ers. Prompt management in- cludes saving prompts with col- lections, metadata in the form of tags and notes, versioning and deletion. For more information, see <a href="#">Generative AI Hub in SAP AI Core Overview [page 60]</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -12-2 0	2023 -12-2 0

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est	Avail- able
										Revi- sion	as of
SAP AI Core	Clou- d Foun- dry	Artifact sig- natures for artifact out- puts from executions	Artifact signatures (hashes) can be generated and made availa- ble to other executions and de- ployments to verify the integrity of an artifact. For more informa- tion, see <a href="#">Using Artifact Signa- tures</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -07-3 1	2023 -07-3 1
SAP AI Core	Clou- d Foun- dry	Template Generator	A wizard to generate workflow and serving templates in VS Code. Using user responses, it simplifies and automates the template writing process.  For more information, see <a href="#">SAP AI Core toolkit documentation</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -08- 04	2023 -08- 04
SAP AI Core	Clou- d Foun- dry	SAP AI Core Toolkit	SAP AI Core is available through the VS Code GUI, through the SAP AI Core toolkit extension.  For more information, see <a href="#">SAP AI Core toolkit documentation</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -08- 04	2023 -08- 04
SAP AI Core	Clou- d Foun- dry	LLM Pack- age	The content package for large language models for SAP AI Core simplifies the deployment of large language models with integrated and automated work- flows.  For more information, see <a href="#">PyPi LLM</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -05- 31	2023 -05- 31
SAP AI Core	Clou- d Foun- dry	Dataset API	You can upload, download, and delete artifacts using the SAP AI Core Dataset API, when direct access to files in the object store is not possible or desirable. Cur- rently, *****Using a Third-Party API Platform***** and curl inter- faces are supported.	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -04- 24	2023 -05- 02

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Metadata in Response to List Exe- cutables	When you use the endpoint to list executables, the re- sponse body now contains metadata about the param- eters and artifacts. For pa- rameters, the <code>description</code> and <code>default</code> values are re- turned. For artifacts, the <code>kind</code> , <code>description</code> , and <code>labels</code> can be added using annotations.  For more information, see <a href="#">List Executables</a> , <a href="#">Workflow Tem- plates</a> , and <a href="#">Serving Templates</a> .	Info only	Gen- eral Avail- abil- ity	Cha- nged	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -04- 02	2023 -04- 02
SAP AI Core	Clou- d Foun- dry	Contribute to Our Doc- umentation	At SAP, we endeavor to make sure that our documentation works for you. If you feel that something is missing or that something doesn't quite hit the mark, you can now provide feed- back and suggest changes di- rectly from SAP Help Portal. You can do so in one of two ways:  <ul style="list-style-type: none"> <li>Click <a href="#">Edit</a> in the toolbar to open the document in Gi- tHub. There, you can sug- gest a change and submit a pull request for us to review.</li> <li>Click <a href="#">Feedback</a> in the tool- bar to create a GitHub is- sue and tell us how we can improve the documentation for you.</li> </ul> For more information, including guidelines on how to contribute, see <a href="#">Open Documentation Initia- tive</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -03-1 3	2023 -03-1 3

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Bulk PATCH Endpoint to STOP or DELETE Multiple Ex- ecutions or Deploy- ments	Executions and Deployments can now receive PATCH requests for bulk adjustments, provided bulkUpdates is enabled in the relevant template. For more information, see <a href="#">Workflow Tem- plates</a> and <a href="#">Servng Templates</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -02- 27	2023 -02- 09
SAP AI Core	Clou- d Foun- dry	Sync End- point for Ar- goCD	In addition to automatically syncing applications, you can re- quest a sync manually by using an API endpoint.  For more information on syncing applications, see <a href="#">Create an Ap- plication [page 91]</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -02- 27	2023 -02- 09
SAP AI Core	Clou- d Foun- dry	WebHDFS Artifacts	WebHDFS artifacts are now sup- ported.  For more information about WebHDFS artifacts on SAP AI Launchpad, see <a href="#">Register an Ob- ject Store Secret [page 102]</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -02- 27	2023 -02- 09
SAP AI Core	Clou- d Foun- dry	Periodic Scheduling	Executions can be run automati- cally, to a prepared schedule. For more information, see <a href="#">Training Schedules</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2023 -02- 27	2023 -02- 09
SAP AI Core	Clou- d Foun- dry	Deploy- ment Dura- tion can be Limited	You can now use the ttl pa- rameter to limit the duration of deployments to hours, days, or weeks.	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -11-2 0	2022 -11-2 0

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est	Avail- able
										Revi- sion	as of
SAP AI Core	Clou- d Foun- dry	The YAML Files for Starter Tu- torials Available Directly from Gi- tHub	The relevant tutorial steps have been updated to include the link to the associated files. The YAML code can also be copied and pasted directly from the tutorials as before.	Info only	Gen- eral Avail- abil- ity	Cha- nged	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -11-2 0	2022 -11-2 0
SAP AI Core	Clou- d Foun- dry	Enhance- ments to the GET Deployme- nts API Call Re- sponse	The GET <code>Deployments</code> re- sponse provides the number of min and max and running replicas, and the resource plan name. For more information, see <a href="#">the SAP AI Core API specifica- tion</a> .	Info only	Gen- eral Avail- abil- ity	Cha- nged	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -10-1 8	2022 -10-1 8
SAP AI Core	Clou- d Foun- dry	Azure Blob Storage Supported	You can now register Azure Blob Storage secrets and use them for Model Serving.  For more information about reg- istering secret, see <a href="#">Register an Object Store Secret [page 102]</a>	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -10-1 8	2022 -10-1 8
SAP AI Core	Clou- d Foun- dry	Free Serv- ice Plan	To try SAP AI Core for free, you can use a free service plan. A free service plan can be easily upgraded to a standard plan, re- taining your users and data.  For more information, see <a href="#">Set Up the Free Plan [page 71]</a> .	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -10-1 8	2022 -10-1 8
SAP AI Core	Clou- d Foun- dry	Deploy- ment Dura- tion Varia- ble	You can now limit the duration of a deployment by specifying the length of time that it should run for, in whole minutes hours or days.	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -10-0 4	2022 -10-0 4

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Metrics Ex- tension	If you have an AI API-ena- bled runtime, you can use the new “Metrics” extension to query which capabilities of the <code>metrics</code> endpoint are sup- ported. You can specify the met- rics capabilities to a very fine level of granularity, which lets you react in your client imple- mentation accordingly.  For more information, see <a href="#">Met- rics Extension</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble		2022 -09-1 9	2022 -09-1 9
SAP AI Core	Clou- d Foun- dry	Git Reposi- tory Name No Longer Mandatory	When you register your git re- pository, you no longer have to enter the repository name. The field has been removed. Reposi- tories already registered with names assigned will still work.	Info only	Gen- eral Avail- abil- ity	An- noun- ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -09-1 9	2022 -09-1 9
SAP AI Core	Clou- d Foun- dry	API Depre- cation and Decommis- sioning	API endpoints <b>POST /lm/ configurations/ {configurationId}/ executions</b> , <b>POST /lm/ configurations/ {configurationId}/ deployments</b> , <b>GET lm/ kpis</b> and <b>GET /analytics/ resourceGroups</b> are deprecated. Please update your existing API calls. For timetables and more informa- tion, see SAP Note <a href="#">3239609</a> .	Re- quire- d	Dep- re- cate- d	Cha- nged	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -09- 05	2022 -09- 05

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	Cha- nged Type	Tech nol- ogy	Not ap- pli- ca- ble	Mod- ular	Lat- est Revi- sion	Avail- able as of
									Line of Busi- ness Proc- ess Product		
SAP AI Core	Clou- d Foun- dry	Serving Component Upgraded to <b>KServe</b> Version 0.7 for Security Compliance	The upgrade means some changes to serving templates: the <code>spec.template.api</code> version changes from <code>servicing.kubeflow.org/ v1beta1</code> to <code>servicing.kserve.io/ v1beta1</code> , the <code>spec.template.spec.pre- dictor.containers.name</code> changes from <code>kfserving- container</code> to <code>kserve- container</code> . There is no sup- port for <code>apiVersion:</code> <code>servicing.kubeflow.org/ v1alpha2</code> .  Please update your existing serving templates by September 30, 2022.	Re- quire d	Gen- eral Avail- abil- ity				SAP Business Technology Platform	2022 -08- 22	2022 -08- 22
SAP AI Core	Clou- d Foun- dry	Supported Object Stores	SAP AI Core supports multiple hyperscaler object stores, such as Amazon S3, OSS, and HANA Data Lake (HDL).  For more information about reg- istering secret, see <a href="#">Register an Object Store Secret [page 102]</a>	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -06- 30	2022 -06- 30
SAP AI Core	Clou- d Foun- dry	Enhanced Serving Template Parameters	The <code>executables.ai.sap.com</code> <code>/cascade-update- deployments</code> parameter can be used in the serving template, to update associated deploy- ments automatically. For more information, <a href="#">Serving Templates</a> and <a href="#">Change Serving Template and Update Deployments</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -06-1 8	2022 -06-1 8

Tech nical Com- po- nent	Envi- ron- men- t	Title	Description	Ac- tion	Life- cy- cle	An- Type	Line of Busi- ness	Mod- ular Busi- ness Proc- ess	Product	Lat- est Revi- sion	Avail- able as of
SAP AI Core	Clou- d Foun- dry	Tracking Service Im- provement	The performance of tracking is improved.	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -06-1 8	2022 -06-1 8
SAP AI Core	Clou- d Foun- dry	Documen- tation for the AI api client SDK has moved to PyPi.org.	Documentation for the AI api cli- ent SDK has moved to PyPi.org. For more information, <a href="#">Libraries and SDKs</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -06- 07	2022 -06- 07
SAP AI Core	Clou- d Foun- dry	Documen- tation for the sap-ai- sdk-core has moved to PyPi.org.	Documentation for the sap-ai- sdk-core has moved to PyPi.org. For more information, <a href="#">Libraries and SDKs</a> .	Info only	Gen- eral Avail- abil- ity	An- noun ce- men- t	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -06- 07	2022 -06- 07
SAP AI Core	Clou- d Foun- dry	Additional Query Pa- rameter for Metrics	The \$select query parameter can be used to retrieve metric resource data, such as metrics or custom info selectively. For more information see, <a href="#">Querying Metric Data</a> .	Info only	Gen- eral Avail- abil- ity	New	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -05- 05	2022 -05- 05
SAP AI Core	Clou- d Foun- dry	Horizontal Scaling for Resource Group Op- erator Ena- bled	The Resource Group Operator now allows horizontal scalability, improving performance.	Info only	Gen- eral Avail- abil- ity	Cha- nged	Tech nol- ogy	Not ap- pli- ca- ble	SAP Business Technology Platform	2022 -04- 09	2022 -04- 09

Technical Component	Environment	Title	Description	Action	Lifecycle	Type	Line of Business	Modular Business Process	Product	Lat-	Avail
										est	able
Revision	of										
SAP AI Core	Cloud Foundry	Resource Group Id Length Extended	The resource group Id length was limited to 10 characters, however longer Ids are now supported. Ids must be of length minimum: 3, maximum: 253. The first and last characters must be either a lower case letter, an upper case letter or, a number. Character entries from the second to penultimate, must be either a lower case letter, an upper case letter, a number, a full stop, or a hyphen. No other special characters are permitted.	Info only	General Availability	Changed	Technology	Not applicable	SAP Business Technology Platform	2022-04-09	2022-04-09
SAP AI Core	Cloud Foundry	Generic Secret Support at Main Tenant Level	You can now store generic secrets at the main tenant level and use them with a generic service broker. For more information, see <a href="#">Register Generic Secrets</a> .	Info only	General Availability	New	Technology	Not applicable	SAP Business Technology Platform	2022-02-19	2022-02-19
SAP AI Core	Cloud Foundry	AI API meta Endpoint	The new meta API endpoint lets clients query the capabilities of a runtime engine that implements AI API. Querying the meta endpoint returns information that the client can use to trigger an appropriate response. For example, SAP AI Launchpad could enable or disable certain features based on the capability of the runtime engine implementation of AI API. For more information, see <a href="#">AI API Runtime Implementations</a> .	Info only	General Availability	New	Technology	Not applicable	SAP Business Technology Platform	2022-02-19	2022-02-19

# 3 Concepts

In this section, we'll explore some of the concepts surrounding both the SAP AI Core and SAP AI Launchpad services.

Term	Definition
AI API	An application programming interface that manages and organizes AI artifacts and workflows (such as training scripts, datasets, models, and model servers) across multiple AI runtimes (environments where AI models are executed, such as cloud platforms or edge devices) to facilitate the development, deployment, and monitoring of AI applications.
AI API connection	A link between SAP AI Launchpad and a runtime. This connection is established through the AI API. For more information, see About the AI API.
AI scenario consumer	A user, either paying or non-paying, who utilizes an AI-powered software application, API, or platform for tasks such as automation, prediction, or decision-making. The consumer may have varying levels of technical expertise, ranging from non-technical end-users to experienced developers.
AI scenario producer	The developer or provider responsible for creating, maintaining, and updating an AI scenario. AI scenarios can be managed in three ways: <ul style="list-style-type: none"><li>• SAP-managed: Developed and maintained by SAP</li><li>• Partner-managed: Developed and maintained by SAP partners</li><li>• In-house: Developed and maintained by the organization using the AI scenario</li></ul>
AI service	A Software as a Service (SaaS) offering provided by SAP Business Technology Platform that enables users to leverage artificial intelligence capabilities for specific use cases. It is typically accessed through an AI API and can be used to expose and deploy AI scenarios as consumable services.
AI use case	A specific application of AI technology designed to generate business value by improving efficiency, reducing costs, or enhancing customer satisfaction. Examples include invoice matching in finance, product recommendations in e-commerce, product review classification in customer service, fraud detection in banking, and predictive maintenance in manufacturing.
artifact	Any tangible by-product produced or consumed by an execution or deployment. Artifacts can include data, files, binaries, libraries, packages, or other resources. They serve as inputs or outputs for various stages of the software lifecycle and are often versioned to track changes and maintain a history of the development process.
configuration	A collection of parameters, artifact references (such as datasets or models), and environment settings that are used to instantiate and run an execution or deployment of an executable or template. A configuration binds specific values to input parameters and specifies the versions of artifacts to be used. It is required to run an execution or deployment, and multiple configurations can be associated with a single executable or template (a 1:n relationship). Configurations are used in both training and serving processes to train models or serve predictions, respectively. They provide a way to customize and control the behavior of the executable or template for different scenarios or use cases.

Term	Definition
dataset	<p>A collection of data used for training, testing, or analysis in artificial intelligence and machine learning applications. A dataset has the following characteristics:</p> <ul style="list-style-type: none"> <li>• References a specific data source, such as a file, database, or API</li> <li>• Includes metadata describing the structure, format, and content of the data</li> <li>• May require specific tools, libraries, or credentials to access and process the data</li> <li>• Can be used as input to machine learning models, algorithms, or data processing pipelines</li> <li>• Is uniquely identified by an ID, which is used to reference the dataset in configurations and bind it to specific components or tasks</li> <li>• May be versioned to track changes and ensure reproducibility of results</li> <li>• Can be divided into subsets for training, validation, and testing purposes</li> <li>• May undergo preprocessing, cleaning, or augmentation steps to prepare the data for use in machine learning tasks</li> </ul>
deployment	<p>An instance of a model serving template (a serving executable or deployment template) that is configured to use a model artifact and apply it to data passed in a serving request (for example, for prediction). A successful deployment creates a model server and generates a deployment URL for inference. As input, a deployment takes one or more models and parameters from a configuration.</p> <p>The serving executable or deployment template defines the expected parameters and input dataset required for the serving process. Values for these parameters and input models are provided by a configuration.</p> <p>Deployments are implemented on a runtime that produces HTTPS endpoints, enabling secure access to the deployed models for inference.</p>
deployment template	<p>A template that defines the parameters and inputs required to serve or deploy one or more trained models. The template is instantiated with a specific configuration, which provides the necessary parameters and inputs. The resulting instance of the template is then used to serve the trained models. Deployment templates are used in the application to facilitate the deployment and serving of machine learning models.</p> <p>A deployment template is also known as a "serving executable".</p>
embedding	<p>A dense, low-dimensional vector representation that captures the semantic and syntactic information of a discrete input item, such as a word, phrase, or other entity, learned from a large dataset. Embeddings are typically used as input features in machine learning models to represent the underlying meaning and relationships of the items in a continuous space.</p>

Term	Definition
executable	<p>A reusable template that defines a workflow or pipeline for tasks such as training a machine learning model or creating a deployment. It contains placeholders for input artifacts (datasets or models) and parameters (custom key-pair values) that enable the template to be reused in different scenarios.</p> <p>There are two types of executables:</p> <ul style="list-style-type: none"> <li>• Non-deployable executables: When instantiated, they result in executions that may produce output artifacts.</li> <li>• Deployable executables: When instantiated, they result in deployments that generate URLs for inference.</li> </ul> <p>Executables are referred to as “templates” within the application. They can have user-defined labels applied to them, which are listed with the executable’s details.</p> <p>The code within an executable defines the training pipeline or model deployment pipeline, leveraging the placeholders for input artifacts and parameters to make the template reusable.</p>
execution	<p>A workflow execution, also known as a “run” in the app, is an instance of a non-deployable executable that represents a single run of a pipeline. There are two types of executions:</p> <ul style="list-style-type: none"> <li>• Training execution: A run of a training pipeline that takes input artifacts and parameters, and produces trained AI models as output artifacts.</li> <li>• Batch-inferencing execution: A run of a batch-inferencing pipeline that takes input artifacts and parameters, and produces result sets as output artifacts.</li> </ul> <p>Each execution is associated with metadata, including metrics, labels, tags, and custom information, which can be queried using the execution ID. For example, a training execution might have metrics like accuracy and loss, while a batch-inferencing execution might have metrics like processing time and number of processed records.</p>
function	<p>A modular, reusable piece of functionality that can be utilized across different scenarios or workflows. Functions are designed to perform specific tasks or computations and can be combined or integrated into larger scenarios to achieve complex business objectives. For example, a data preprocessing function can be used to clean and transform input data before feeding it into a machine learning model, while a model evaluation function can be used to assess the performance of a trained model. Functions promote code reusability, maintainability, and efficiency in building AI solutions.</p>
input artifact	<p>Placeholder in an executable or template that enables the attachment of datasets or models required for the execution of an AI workflow or pipeline. These artifacts can include training data, validation data, pre-trained models, or any other data or model assets needed for the AI process.</p>
job executable	<p>A simplified representation of a workflow executable. A job executable encapsulates the essential components and logic of a workflow, allowing users to easily understand and manage the execution of AI and machine learning tasks within the Functions Explorer interface of SAP AI Launchpad.</p>
job template	<p>A predefined configuration that specifies the parameters and resources required to execute a long-running AI process, such as model training, batch inference, or data preprocessing. It defines the Docker image, compute resources, environment variables, and other settings needed to run the job. Job templates act as blueprints that can be instantiated to create and execute specific job instances with customized parameters.</p>

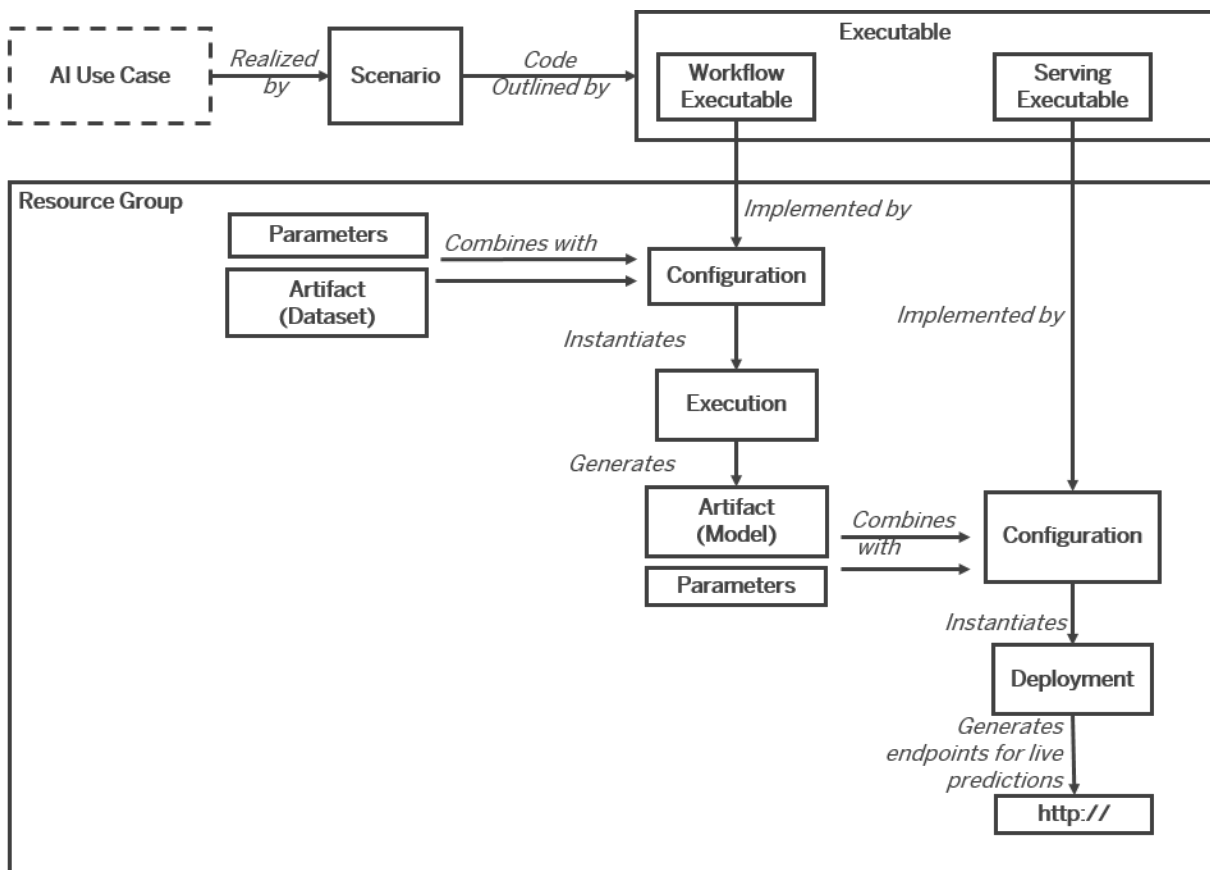
Term	Definition
knowledge graph	A structured representation of entities, their attributes, and the relationships between them, which can be inferred from various data sources, including structured, semi-structured, and unstructured data. Knowledge graphs enable machines to understand and reason about complex real-world concepts and their interconnections, facilitating tasks such as semantic search, recommendation systems, and question answering. Examples of knowledge graphs include ontologies, taxonomies, and knowledge bases.
label	A key-value pair attached to a metric to provide additional context and metadata. Labels are used to classify and categorize metrics, enabling users to filter, group, and aggregate data for better insights and analysis. A set of labels can be applied to each instance of a metric record, allowing for more granular and targeted monitoring.
metrics	A key/value pair where the value is numeric. Metrics are used to measure and monitor the performance, progress, and quality of a model during the training process. Common examples of metrics include accuracy, precision, recall, F1 score, and mean squared error. Metrics can have optional step, timestamp, and label fields, which provide additional information about the metric's context. Every metric (and associated labels), tag, and custom info must be associated with a specific training execution. This association allows for proper organization, tracking, and comparison of different training runs. Once the metric, tag, and custom info are saved, they can be queried using an execution ID. SAP AI Launchpad provides a user-friendly interface to visualize and analyze the captured metrics, enabling users to monitor the model's performance and make informed decisions based on the data.
model	<p>An artifact that is the output of a machine learning training process, representing the learned patterns, parameters, and structure of a trained system.</p> <ul style="list-style-type: none"> <li>• A model is generated by a training process, which optimizes the model's parameters based on training data.</li> <li>• A model consists of one or more files stored in a hyperscaler storage system (such as SAP AI Core's connected data storage) or the data lake for SAP HANA Cloud.</li> <li>• Each model is uniquely identified by a model ID, which is used in a configuration to bind the model as an input artifact to the serving executable for deployment.</li> <li>• Models can be manually uploaded to the connected data storage or automatically generated and uploaded as the output of a training execution in SAP AI Launchpad.</li> <li>• The model encapsulates the learned knowledge of the AI system and is used for making predictions or decisions on new, unseen data during inference.</li> <li>• Different types of models exist, such as neural networks, decision trees, or clustering models, each suited for different AI tasks and data types.</li> <li>• In SAP AI Core, models are managed and versioned artifacts that can be deployed to serving endpoints for production inference.</li> </ul>
model serving template	A predefined configuration that specifies how a trained machine learning model is to be deployed and served for inference in a production environment, typically including details such as the model framework, runtime environment, resources required, and API endpoints.
operations	All activities within the AI lifecycle, including data preparation, model training, model deployment, application integration, model monitoring, and continuous improvement.

Term	Definition
output artifact	Generated results, typically AI models, produced by executing or running a training process or pipeline. When AI models are generated from an execution, they are automatically uploaded and registered to the connected data lake for SAP HANA Cloud or to the hyperscaler data storage, which is a scalable cloud storage solution provided by major cloud service providers such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform (GCP).
parameter	A placeholder in an executable (script, program, or model) to which a value of a specific data type (String, Integer, Float, Boolean, List, or Dictionary) is assigned during runtime. The values are provided by configurations, which are settings or files that define the parameter values. Parameters are used to provide input data, settings, or options to the executable for customizing its behavior in executions or deployments.
prompt	A natural language instruction or query given to a generative AI model to elicit a response. An AI platform like SAP AI Launchpad includes capabilities for prompt experimentation, management, and administration: <ul style="list-style-type: none"> <li>• Prompt experimentation: Creating and running natural language prompts with a choice of large language models and adjustable parameters to test and optimize prompts.</li> <li>• Prompt management: Saving, organizing, and managing prompts using collections, metadata (tags and notes), versioning to track changes over time, and deletion when no longer needed.</li> <li>• Prompt administration: Tools for governing prompt usage, monitoring prompt performance, controlling access and permissions, and ensuring responsible AI practices.</li> </ul>
resource group	A dedicated workspace for a specific AI scenario or use case. It allows users to organize, manage, and collaborate on related AI entities such as configurations, executions, deployments, and artifacts within a defined scope. Resource Groups ensure isolation and access control, enabling users to work on their AI projects independently while sharing common resources and artifacts as needed. Each Resource Group is associated with a unique subscription to an SAP AI Launchpad tenant.
result set	A data artifact or dataset that contains the results of a batch inference run (execution). A batch inference run uses a trained machine learning model to process an input dataset containing data points or instances and generates predictions or inferences for each data point. The result set is an output artifact that stores these predictions or inferences.
run	A training process that generates a model or models based on a run template. Also known as an "execution".
run template	A template that defines the components required for an AI pipeline within the ML Operations application. It specifies placeholders for parameters, input artifacts (such as datasets), and output artifacts (such as trained models) that are necessary to instantiate and run the executable successfully. The actual values for these placeholders are provided by a configuration, which is a set of settings and parameters specific to a particular use case. In SAP AI Launchpad, a run template is also referred to as a "workflow executable".
runtime	The environment or platform that provides the necessary processing resources to execute AI and machine learning workloads, such as training and inference. It enables businesses to make their applications intelligent by leveraging AI and machine learning technologies, and allows them to train AI services using their data to automate tasks and processes. SAP AI Core is an example of a runtime.

Term	Definition
SAP AI Core	<p>A service in the SAP Business Technology Platform designed to handle the execution and operations of your AI assets (machine learning models, datasets, and other AI-related resources) in a standardized, scalable, and hyperscaler-agnostic way, meaning it can work with different cloud service providers such as AWS, Azure, and Google Cloud Platform. SAP AI Core provides seamless integration with your SAP solutions, enabling you to easily incorporate AI capabilities into applications like SAP S/4HANA, SAP SuccessFactors, or SAP Customer Experience. Any AI function can be realized using popular open-source frameworks such as TensorFlow, PyTorch, or scikit-learn. SAP AI Core supports the full lifecycle management of AI scenarios, including data preparation, model training, deployment, monitoring, and retraining.</p>
SAP AI Launchpad	<p>A multitenant software as a service (SaaS) application on SAP Business Technology Platform (SAP BTP). Customers and partners can use SAP AI Launchpad as a centralized platform to manage AI use cases (scenarios), which are specific applications or problems that can be solved using AI, across multiple instances of AI runtimes (such as SAP AI Core). SAP AI Launchpad also provides access to the generative AI hub, which offers a set of pre-built generative AI models and tools that users can leverage to create and deploy their own generative AI applications.</p>
scenario	<p>An implementation of a specific AI use case within a user's tenant. It consists of a pre-defined set of AI capabilities in the form of executables (pre-built AI models or services that can be directly deployed) and templates (outlines of AI models or services that need to be trained with customer data).</p> <p>A scenario can have multiple versions that correspond to different versions of its contained executables and templates. Placeholders within the executables and templates are populated with customer-specific values using a configuration.</p> <p>The purpose of a scenario is to group executables and templates related to an AI use case and make them available to all AI consumers in the tenant, while also enabling version tracking as the AI models and services evolve over time.</p>
serving executable	<p>A predefined AI pipeline that encapsulates the necessary components and logic to deploy an AI model or application. It serves as a blueprint for creating a specific deployment instance.</p> <p>A deployment template defines placeholders for parameters, input artifacts (such as datasets), and output artifacts (such as trained models). These placeholders are filled with actual values provided by a configuration when the template is instantiated.</p> <p>Instantiating a deployment template involves providing the required configuration, which specifies the concrete values for the parameters and input artifacts. This process creates a deployment instance that can be executed to train, evaluate, or serve an AI model or application.</p> <p>In the SAP AI Launchpad, deployment templates are referred to as "serving executables".</p>
tag	<p>A name/value pair that is used to categorize and organize test executions. Tags allow for the segregation and grouping of test executions based on specific criteria, making it easier to manage, analyze, and report on the results. For example, you can assign tags to a group of selected test executions to indicate their purpose, priority, or other relevant characteristics. A set of tags can be associated with a MetricResource, which is an entity that represents a specific metric or measurement. The MetricResource, in turn, is linked to an execution, establishing a connection between the tags and the corresponding test run. This relationship enables the effective tracking, monitoring, and evaluation of test executions based on their assigned tags.</p>

Term	Definition
tenant	A logically separated customer instance that represents an organization or a company. Each tenant has its own isolated collection of customized content and services, which are available only to that specific tenant, its users, and the service provider managing the multi-tenant environment.
training	The process of running a machine learning algorithm on a dataset to produce a trained model. The trained model is an artifact that captures the patterns and relationships learned from the data, which can then be used to make predictions or decisions on new, unseen data.
workflow executable	A template that defines the components required for an AI pipeline within the ML Operations application. It specifies placeholders for parameters, input artifacts (such as datasets), and output artifacts (such as trained models) that are necessary to instantiate and run the executable successfully. The actual values for these placeholders are provided by a configuration, which is a set of settings and parameters specific to a particular use case. In SAP AI Launchpad, a workflow executable is also referred to as a "run template".

You can see how these concepts interact from the following diagram:



## 3.1 AI API Overview

The AI API lets you manage your AI assets (such as training scripts, data, models, and model servers) across multiple runtimes.

Argo workflows and serving templates, as well as their execution and deployment, are managed using the SAP AI Core implementation of the AI API. In SAP AI Core, the Argo workflow and serving templates are mapped under the concept of `Executable`. For the mapping mechanism to work, the Argo workflows and serving templates require certain attributes in the metadata section of the `YAML` file. These attributes are shared by both template types.

SAP AI Core provides additional APIs that are runtime-specific. These are available in the AI Core API specification, which is an extension of the AI API specification.

### Related Information

[AI Core API](#) 

[AI API](#)

## AI API Runtime Implementations

The AI API specification is a general specification for the lifecycle management of machine learning artifacts. SAP AI Core is one specific runtime implementation of the AI API specification. It is also possible to provide other runtime implementations of the AI API specification, independent of SAP AI Core. This section describes the necessary boundary conditions and implementation requirements.

The benefit of using AI API is that clients can integrate with all AI API-enabled runtime implementations. For example, SAP AI Launchpad can interact with custom runtime implementations as long as the same APIs are provided. Intelligent Scenario Lifecycle Management can also integrate with AI API-enabled runtimes. The SAP AI SDK Base (Python) can also be used (for more information, see [sap-ai-sdk-core](#)).

### AI API Specification

The AI API specification comprises the following parts:

- A main specification
- Extensions:
  - Analytics extension
  - Resource group extension
  - Dataset management extension
    - Metrics extension

## → Recommendation

Implement at least the main specification, and then implement the extension specifications based on your use case.

## AI API Runtime Capabilities Endpoint

Meta API is part of the AI API specification (endpoint `/1m/meta`). The implementation must return a configuration response that specifies the capabilities of the AI API runtime implementation.

Meta API allows AI API clients to query the capabilities of an AI API implementation so that they can select which commands or user interfaces are available. For example, some AI API runtimes may offer executions but not deployments. They may also offer logs for executions and not for deployments. As an example, if a client of SAP AI Core such as SAP AI Launchpad queries the Meta API endpoint of SAP AI Core, the response will be for example:

```
{
  "aiApi": {
    "capabilities": {
      "logs": {
        "deployments": true,
        "executions": true
      },
      "multitenant": true,
      "shareable": true,
      "staticDeployments": true,
      "timeToLiveDeployments": true,
      "userDeployments": true,
      "userExecutions": true,
      "executionSchedules": true
    },
    "limits": {
      "deployments": {
        "maxRunningCount": -1
      },
      "executions": {
        "maxRunningCount": -1
      },
      "minimumFrequencyHour": 1,
      "timeToLiveDeployments": {
        "minimum": "10m",
        "maximum": -1
      }
    },
    "version": "2.18.0"
  },
  "extensions": {
    "analytics": {
      "version": "1.0.0"
    },
    "metrics": {
      "capabilities": {
        "extendedResults": true
      },
      "version": "1.0.0"
    },
    "resourceGroups": {
      "version": "1.2.0"
    }
  }
}
```

```

    },
    "runtimeApiVersion": "2.21.0",
    "runtimeIdentifier": "aicore"
  }

```

SAP AI Launchpad and other clients can then react accordingly and hide the deployments on the user interface for this runtime implementation of AI API.

Capabilities include the following:

Capability	When true, allows users to:
logs.executions	View logs for an execution
logs.deployments	View logs for a deployment
multitenant	Use SAP AI Launchpad as a main tenant user (supports resource groups)
shareable	Clients can share one instance
staticDeployments	Static, always running endpoints for inference are available without the user having to start a deployment
userDeployments	Stop, update, or delete a deployment
userExecutions	Stop or delete an execution
timeToLiveDeployments	The runtime engine allows defining the time until a deployment is automatically deleted
analytics	Review summary information for all tenants
bulkUpdates	Stop or delete up to 100 executions or deployments at once
executionSchedules	Create schedules

Limits include the following:

Limit	Details
deployments.maxRunningCount	Limits the number of running concurrent deployments in a resource group, if any
executions.maxRunningCount	Limits the number of running concurrent executions in a resource group, if any
timeToLiveDeployments.minimum	The minimum possible value for the ttl parameter in a deployment, if supported
timeToLiveDeployments.maximum	The maximum possible value for the ttl parameter in a deployment, if supported
minimumFrequencyHour	The minimum possible value for schedule of an execution, if supported

In addition to the general AI API specification, there is also a number of extensions that cover additional use cases. These might not be implemented in all runtime engines.

The extensions are:

Extension	Details
analytics	The analytics extension contains endpoints for fetching analytical information of a resource group or tenant
metrics	The metrics extension contains endpoints for writing to and reading from metrics endpoints, to store and retrieve metrics generated during executions
resourceGroups	The resource group extension contains endpoints for managing resource groups
dataset	The dataset extension contains endpoints for uploading and downloading files

## Related Information

[Serving Templates](#)

[Workflow Templates](#)

[AI API Specification](#)

[Custom Runtime Capabilities Using the Meta API](#)

[Analytics Extension](#)

[Resource Groups Extension](#)

[Intelligent Scenario Lifecycle Management](#)

## 3.2 Resource Groups

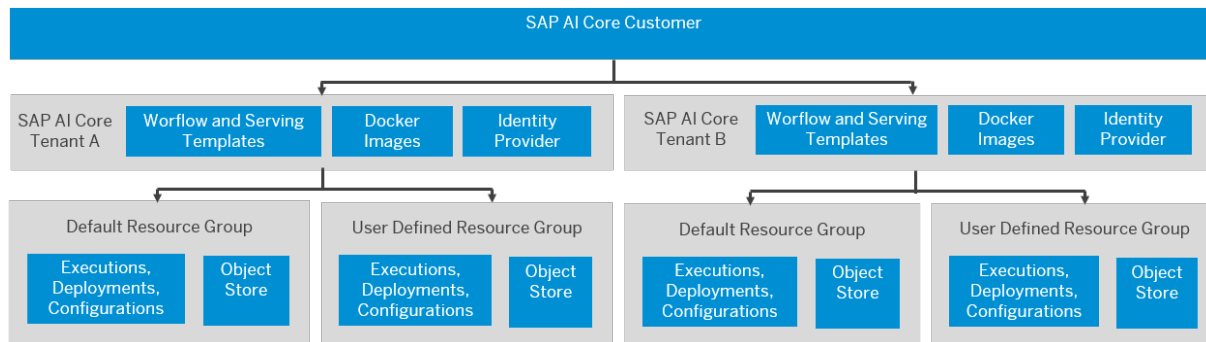
Tenants use resource groups to isolate related ML resources and workloads. Scenarios, executables, and Docker registry secrets are shared across all resource groups.

Resource groups represent a virtual collection of related resources within the scope of one tenant. When a tenant is onboarded, the system immediately creates a default resource group. Tenant administrators can create or delete additional resource groups using the AI API. Tenants can map resource groups based on corresponding usage scenarios.

If a tenant uses resource groups to isolate scenario consumer tenants and those resource groups are deleted, the scenario consumers are deprovisioned. The service does not recognize the scenario consumer of the tenant. The standard XSUAA multitenancy model is followed.

## 3.2.1 Scope of Resources

Resources that are available for tenants and resource groups differ based on the available scope.



### Tenant-Level Resources

Tenant-level resources include:

- Workflow templates
- Serving templates
- Docker registry (containing the Docker images)
- User authentication and authorization (UAA)

User authentication and authorization is based on the SAP AI Core tenant. The tenant is the holder of the access token obtained using the SAP AI Core service key. The SAP AI Core tenant can set the resource group in the request header at runtime, or during lifecycle management, using the AI API. If the resource group is not set, the default resource group is used.

### Resource Group-Level Resources

Executables at tenant level are shared across all of the resource groups. At resource group level, the object store is registered by setting the resource group header.

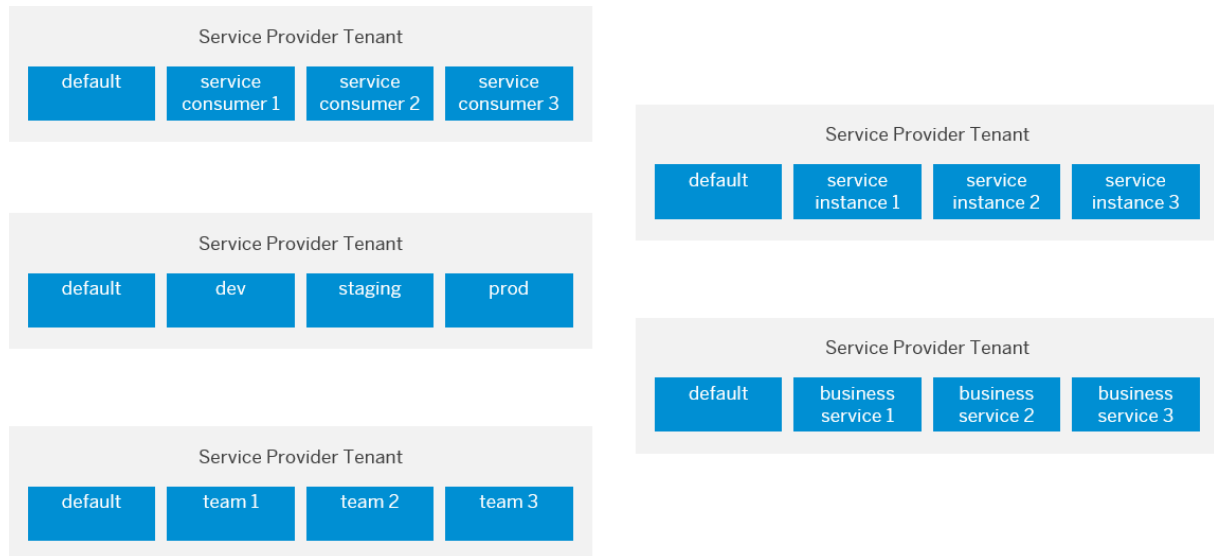
SAP AI Core tenants must consider security aspects in the design of AI functions.

#### → Recommendation

Do not use the same object store bucket with the same AWS IAM user for multiple resource groups.

Runtime entities such as executions, deployments, configurations, and artifacts belong to specific resource groups and cannot be shared across resource groups.

## Examples of Resource Group Mapping



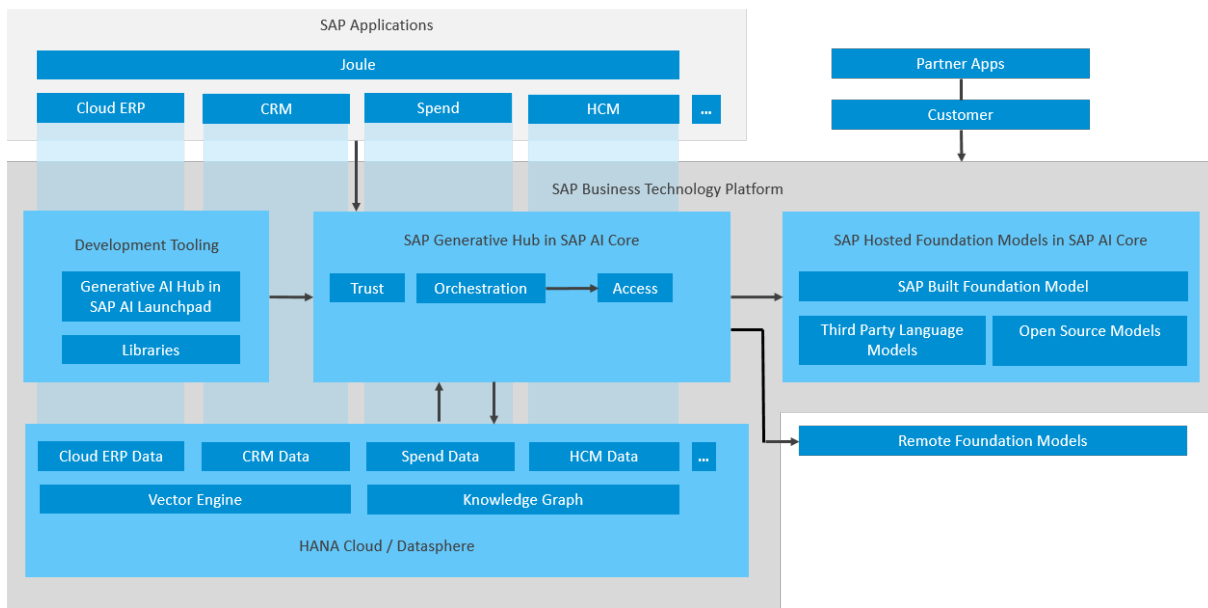
### 3.3 Generative AI Hub in SAP AI Core Overview

The generative AI hub incorporates generative AI into your AI activities in SAP AI Core and SAP AI Launchpad.

Generative AI models are self-supervised, deep learning models trained on vast amounts of unlabeled data. They use AI technology and industrial-scale computational resources to learn complex patterns and semantic knowledge bases. These models excel in tasks like natural language processing (NLP). By parsing inputs, such as prompts, and predicting target words, they return contextually relevant responses in natural language. A single model can handle multiple tasks by using different input formats and output modes.

Generative AI models are general by design, but you can fine-tune them with additional embeddings. In this way, you can make them suitable for specialized or domain-specific use cases.

SAP AI Core and the generative AI hub help you to integrate LLMs and AI into new business processes in a cost-efficient manner.



Generative AI Hub Architecture Overview

# 4 Initial Setup

Get started with SAP AI Core using the standard procedures for the SAP BTP, Cloud Foundry environment or Kyma environment.

[Enabling the Service in Cloud Foundry \[page 62\]](#)

Enable SAP AI Core using the standard procedures for the SAP BTP, Cloud Foundry environment.

[Enabling the Service in the Kyma Environment \[page 83\]](#)

Enable SAP AI Core using the standard procedures for the SAP BTP Kyma environment.

## 4.1 Enabling the Service in Cloud Foundry

Enable SAP AI Core using the standard procedures for the SAP BTP, Cloud Foundry environment.

### → Tip

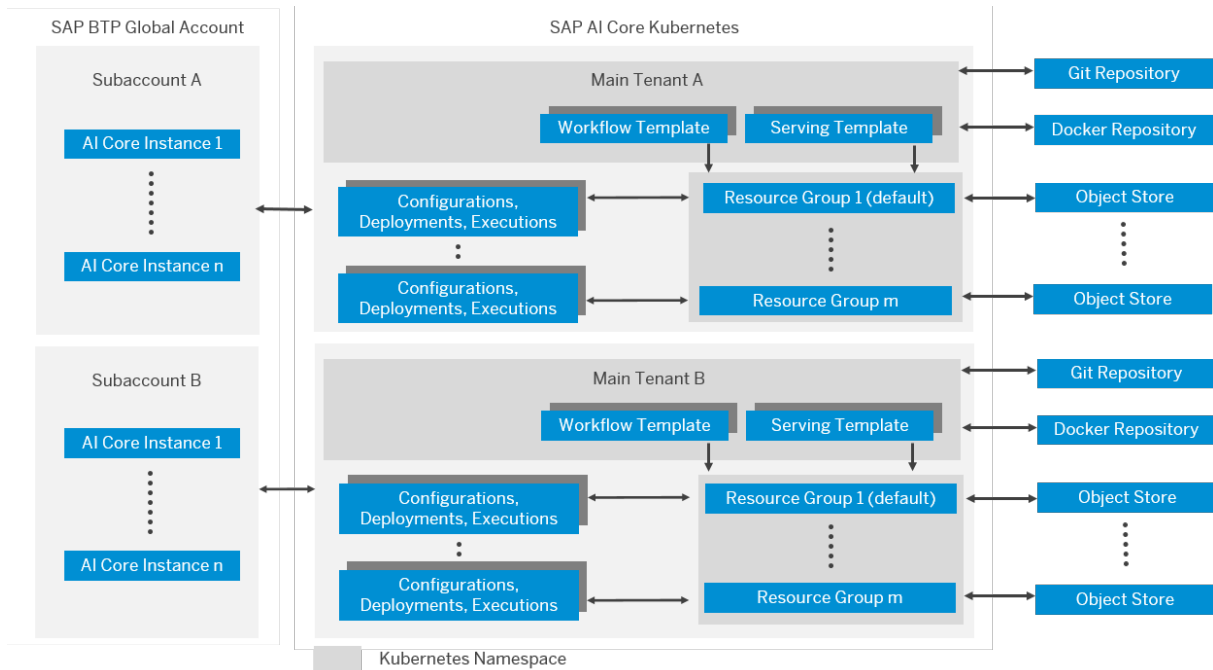
You can use the booster [Set Up Account for SAP AI Core](#) to automate the steps described in this section on the SAP BTP cockpit. For more information, see [Use Boosters for Free Plan Use of SAP AI Core and SAP AI Launchpad](#).

When you provision SAP AI Core from the SAP BTP cockpit in SAP Business Technology Platform, the system generates a service key. This key contains the URLs and credentials you need to access the SAP AI Core instance.

SAP AI Core is a tenant-aware reuse service that isolates tenants based on the zone ID, which represents the subaccount. When you create an SAP AI Core service instance within a subaccount, it represents an SAP AI Core tenant.

### ⓘ Note

The SAP AI Core service doesn't isolate tenants based on the service instance ID. If you create multiple service instances within the same subaccount, they all reference the same SAP AI Core tenant.



Parent topic: [Initial Setup \[page 62\]](#)

## Related Information

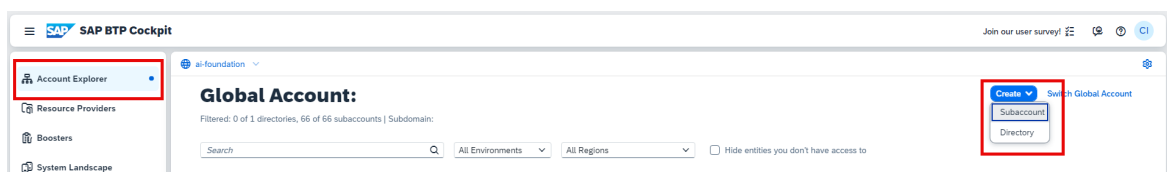
[Enabling the Service in the Kyma Environment \[page 83\]](#)

### 4.1.1 Create a Subaccount

Create a subaccount in your global account using the SAP BTP cockpit.

#### Procedure

1. In the SAP BTP cockpit, choose *Account Explorer* and then click **Create** **Subaccount**.



2. In the *Create Subaccount* dialog, enter a name for your subaccount and select the region. The parent defaults to the name of your global account.

### Create Subaccount

Display Name \*

Region \*

Description

Subdomain \*

Parent \*

> **Advanced**

**Create** **Cancel**

- Optional:** If your subaccount is used for production purposes, under *Advanced* select the *Used for production* checkbox.

This setting does not change the configuration of your subaccount. It is intended to help you manage the production subaccounts in your global account. For example, your cloud operator can refer to it when handling incidents related to mission-critical accounts.

Advanced

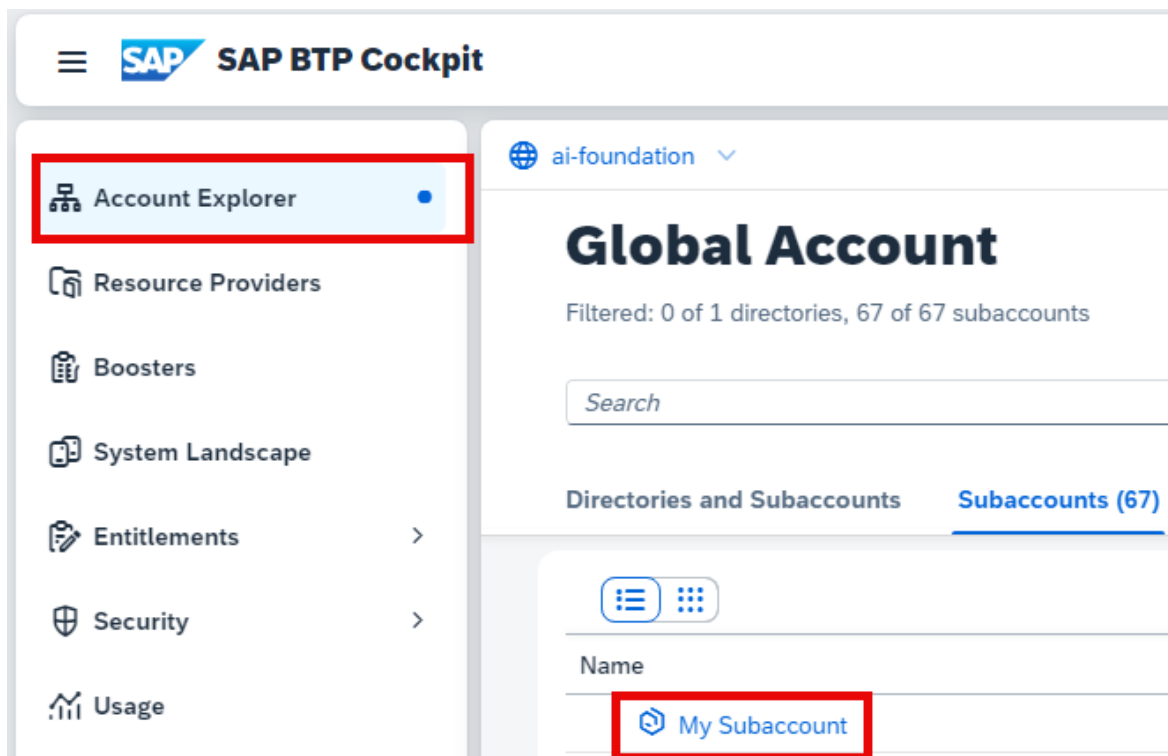
**Used for production** ⓘ

Enable beta features ⓘ

Labels ⓘ

Name:  Values:

- Click *Create*.
- Return to the *Account Explorer* to view your subaccount.



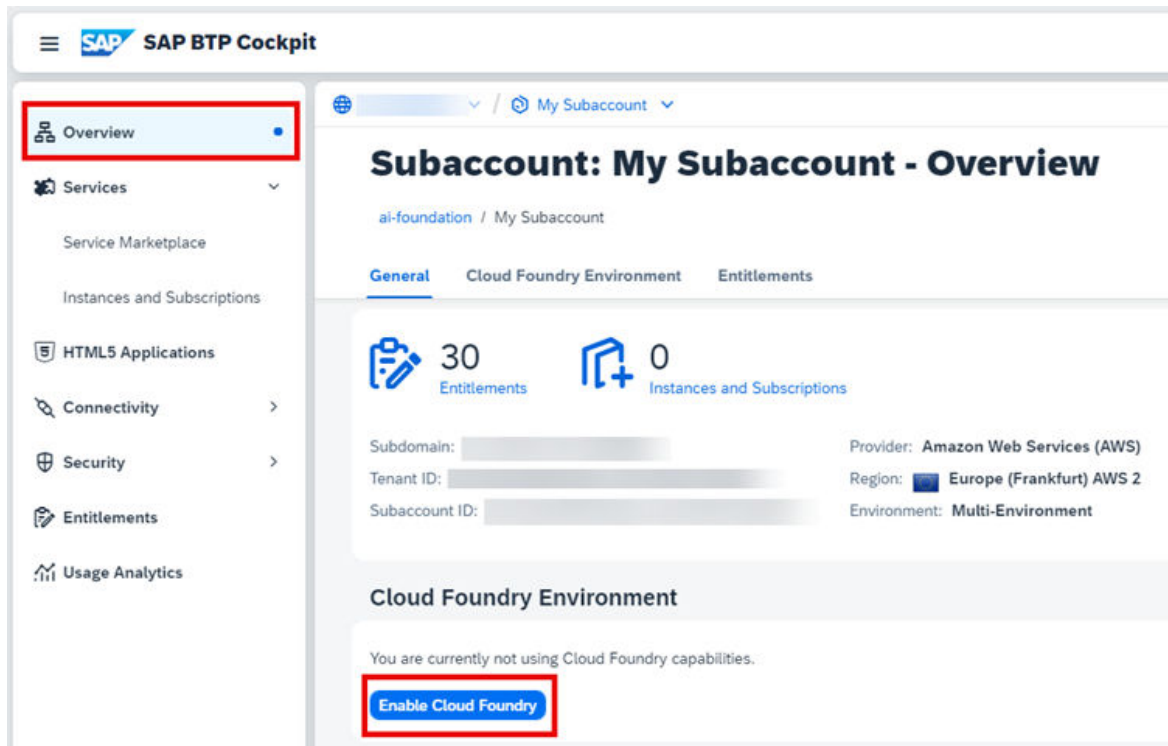
## Related Information

[SAP BTP Cockpit](#)

## 4.1.2 Enable Cloud Foundry

### Procedure

1. Click your subaccount and on the *Overview* page, choose *Enable Cloud Foundry*.

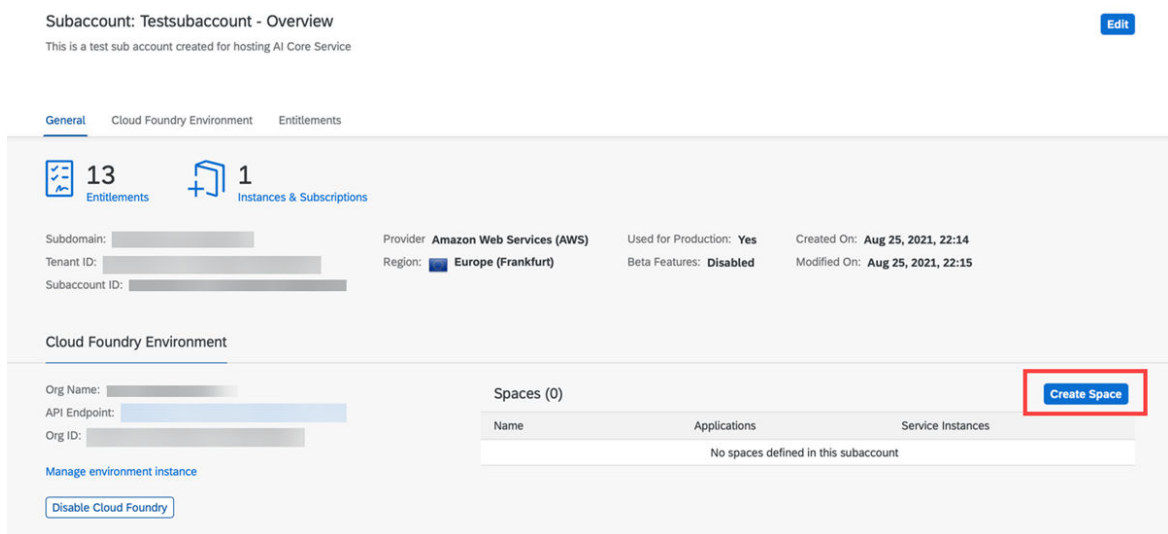


2. Enter the basic information for your Cloud Foundry environment instance and click *Create*.

## 4.1.3 Create a Space

### Procedure

1. On the overview page for your subaccount, choose *Create Space*.



2. Enter a name for your space, assign the required roles, and click *Create*.

**Create Space**

Space Name: \*

Test

Assign space roles to

- Space Manager
- Space Developer
- Space Auditor

**Create** Cancel

## Related Information

[About Roles in the Cloud Foundry Environment](#)

### 4.1.4 Add a Service Plan

Configure the required entitlements to make SAP AI Core accessible in your subaccount.

## Procedure

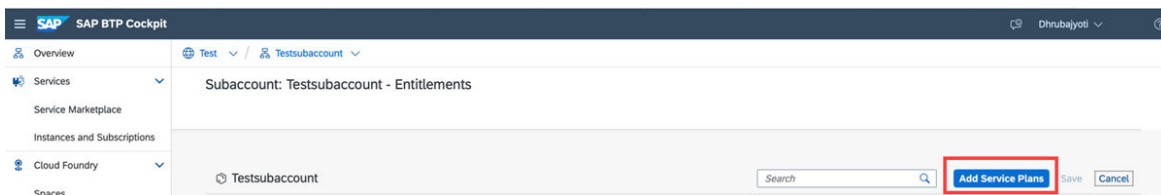
1. In SAP BTP cockpit, navigate to your global account and choose *Entitlements* and then *Entity Assignments*.
2. Use the input help to select your subaccount and click *Edit*.
3. Choose *Edit*.

**Global Account**

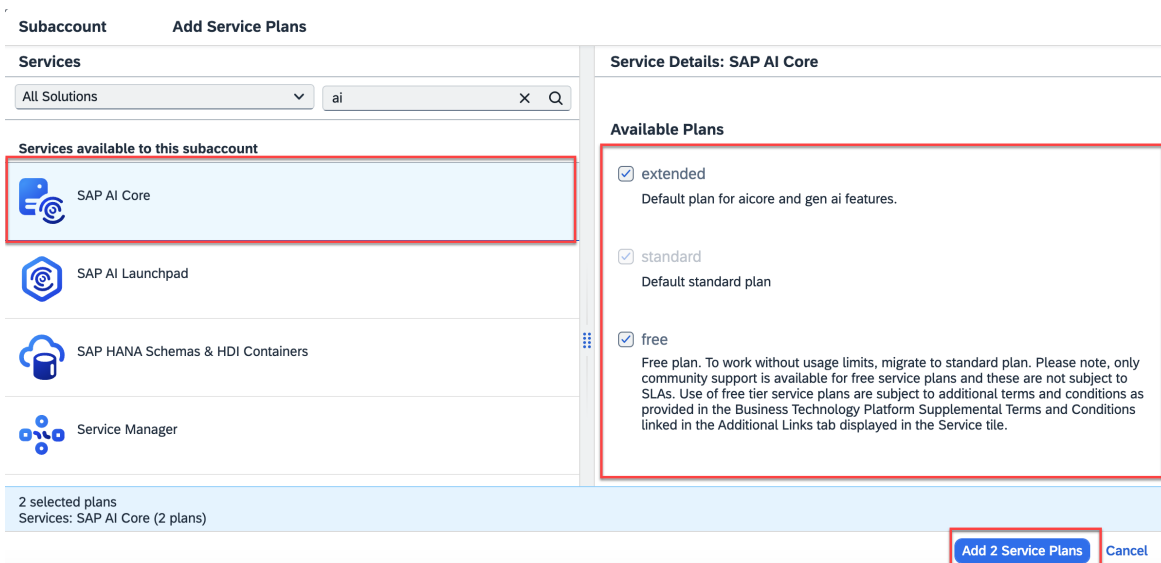
Subaccounts/Directories: My Subaccount

Service	Service Technical Name	Plan	Set Quota Limit	Quota Assignment	Remaining Quota	Actions
[DEPRECATED] Audit Log Retrieval	auditlog-api	default		Assigned (1 unit)		
Application Autoscaler	autoscaler	standard		Assigned (1 unit)		

4. Choose *Add Service Plans*.



5. Select SAP AI Core and pick your chosen service plan or plans.



**Note**  
 To use generative AI capabilities, choose the extended plan. For more information, see [Service Plans \[page 69\]](#).

6. Save your changes.

The screenshot shows a table titled 'Testsubaccount (Unsaved changes)'. The table has columns for Service, Plan, Assign Quota, Subaccount Assignment, Remaining Global Quota, and Actions. The 'Save' button at the top right is highlighted with a red box.

Service	Plan	Assign Quota	Subaccount Assignment	Remaining Global Quota	Actions
Audit Log Viewer	free (Application)	<input type="checkbox"/>	1 shared units	1 shared units	
Cloud Identity Services	application	<input type="checkbox"/>	1 shared units	1 shared units	
Cloud Integration Automation Service	standard (Application)	<input type="checkbox"/>	1 shared units	1 shared units	
	oauth2	<input type="checkbox"/>	1 shared units	1 shared units	
Content Agent Service	application	<input type="checkbox"/>	1 shared units	1 shared units	
HTML5 Application Repository Service	app-host	<input type="checkbox"/>	1 shared units	1 shared units	
	app-runtime	<input type="checkbox"/>	1 shared units	1 shared units	
Master Data Integration (Orchestration)	standard (Application)	<input type="checkbox"/>	1 shared units	1 shared units	

## 4.1.4.1 Service Plans

The SAP AI Core service plan you choose determines pricing, conditions of use, resources, available services, and hosts.

Your choice depends on your use case:

- Use the *Free* plan to explore the service with limited resources and community-only support.
- Use the *Standard* plan for production workloads without generative AI.
- Use the *Extended* plan for production workloads with generative AI hub access.

### → Tip

If you want a guided, hands-on environment, you can explore the generative AI hub using the 30-day free trial. For more information, see [Try now: 30-Day Basic Trial](#).

## Comparison of Service Plans

Feature / Plan	Free (Exploration)	Standard (Production)	Extended (Production + Generative AI)
Account type	Enterprise account (SAP BTP free tier only; not available in SAP BTP trial)	Enterprise account	Enterprise account
Support	Community support only (no SLA)	Full SAP support with SLA	Full SAP support with SLA
Generative AI hub	✗ Not included	✗ Not included	✓ Included
Pricing	Free	Pay per resource + baseline charge	Pay per resource + model/token usage
Instances per subaccount	1 instance (mutually exclusive with Standard)	1+ instances (mutually exclusive with Free)	1+ instances
Executions/Deployments	1 running execution or deployment at a time	Multiple	Multiple
Resource groups	Default group only	Multiple (quota applies)	Multiple (quota applies)
Resource plan	Starter plan only	Choice of resource plans	Choice of resource plans
Upgrade/downgrade rules	Upgrade from Free to Standard is possible. Downgrade from Standard to Free is not possible.	Cannot be created if a Free plan is active in the same subaccount.	Cannot be created if a Free plan is active in the same subaccount.

### Note

You can only run either a Free or a Standard plan in the same subaccount — not both. Extended follows the same rules as Standard.

For supported regions, see [SAP Discovery Center](#) .

## Quotas

### Deployment Quotas

Each tenant is assigned a default quota that limits the number of deployments and replicas per deployment. If you reach this quota, your deployment will not be created and you will be notified accordingly. You can free up your quota by deleting existing deployments.

Alternatively, you can request an increase to your quota by creating a ticket on component **CA-ML-AIC**. Enter the description **Request to Increase Quota** and include details about the size of your increase, whether you want to include deployments, replicas, or both, and your subaccount ID.

### Resource Group Quotas

#### Restriction

The maximum number of resource groups is limited at tenant level to 50. If you reach this limit, you receive an error message. To free up space, delete some resource groups. Alternatively, raise a ticket to increase your quota.

For more information, see [Delete a Resource Group \[page 100\]](#).

For more information, see [Delete a Resource Group](#).

### Tenant-Wide Generic Secrets Quotas

Each tenant can have a maximum of five tenant-wide secrets. If you reach this limit, you receive an error message. To free up space, delete tenant-wide secrets as described at [Delete a Generic Secret \[page 128\]](#). Alternatively, submit a ticket to request an increase in your quota.

## Related Information

[Set Up the Free Plan \[page 71\]](#)

[SAP Discovery Center](#) 

[SAP BTP Service Description Guide](#) 

[Choose an Instance](#)

## 4.1.4.1.1 Set Up the Free Plan

The free plan lets you try out SAP AI Core for testing and familiarization purposes at no cost.

### Prerequisites

You have a global account in the free tier model for SAP BTP (not available in SAP BTP Trial).

### Context

Alternatively, you can use a booster to set up the free plan for SAP AI Core. For more information, see the tutorial [Use Boosters for Free Plan Use of SAP AI Core and SAP AI Launchpad](#).

### Procedure

1. Open your global account in the SAP BTP cockpit.
2. Go to your subaccount.
3. In the navigation area, choose *Instances and Subscriptions*.
4. Choose *New Instance or Subscription*.
5. In the *Service* field, search for SAP AI Core.
6. In the *Plan* field, choose *Free*.

**New Instance or Subscription**

1 Basic Info      2 Parameters      3 Review

Enter basic info for your instance or subscription.

Service: \* ⓘ  
SAP AI Core

Plan: \*  
standard Instance  
free Instance

## Results

A Free plan tenant is created in your subaccount.

## Related Information

[Service Plans \[page 69\]](#)

[Using Free Service Plans](#)

[Getting a Global Account](#)

### 4.1.4.1.2 Update a Service Plan

Update your SAP AI Core service instance from the free plan to a standard or extended plan while keeping your data and models.

## Context

During the update, all metadata and transaction data, including trained models, is retained.

You can also update from the standard plan to the extended plan.

### ⚠ Restriction

You cannot downgrade from the standard or extended plan to the free plan.

Downgrading from the extended plan to the standard plan is also not supported.

If a standard or extended instance is deleted, you cannot create a new standard plan instance.

## Procedure

1. Open your global account in the SAP BTP cockpit.
2. Go to your subaccount.
3. In the navigation area, choose *Instances and Subscriptions*.
4. Search for SAP AI Core.
5. At the end of the subscription row, select the ellipsis (...) and choose *Update*.

free-plan-instance	SAP AI Core	standard	Created	Update
			Created	Create Binding
			Created	Create Service Key
			Created	Delete
			Created	

- In the wizard that opens, select *default* and click *Update Subscription*.

## Results

Free plan restrictions no longer apply.

All data from your free plan is migrated automatically to your new plan.

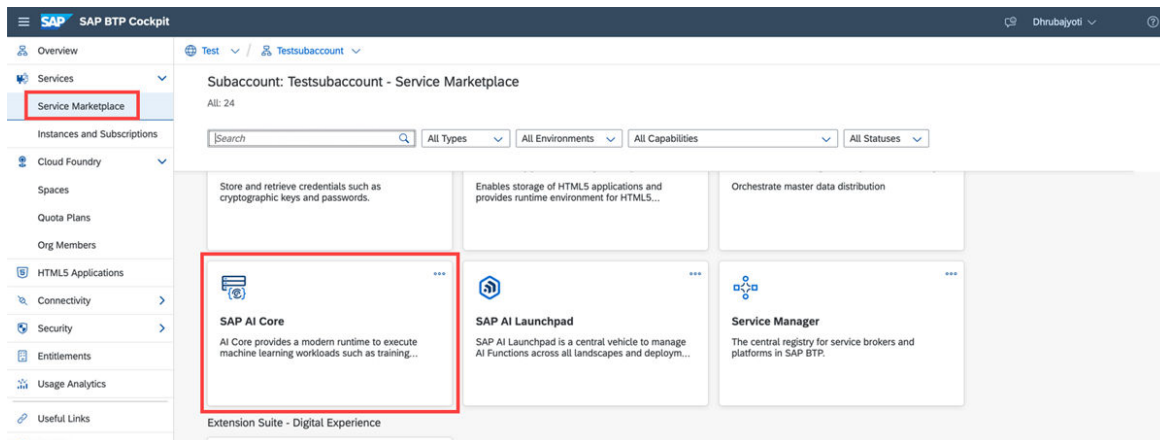
User permissions remain unchanged.

## 4.1.5 Create a Service Instance

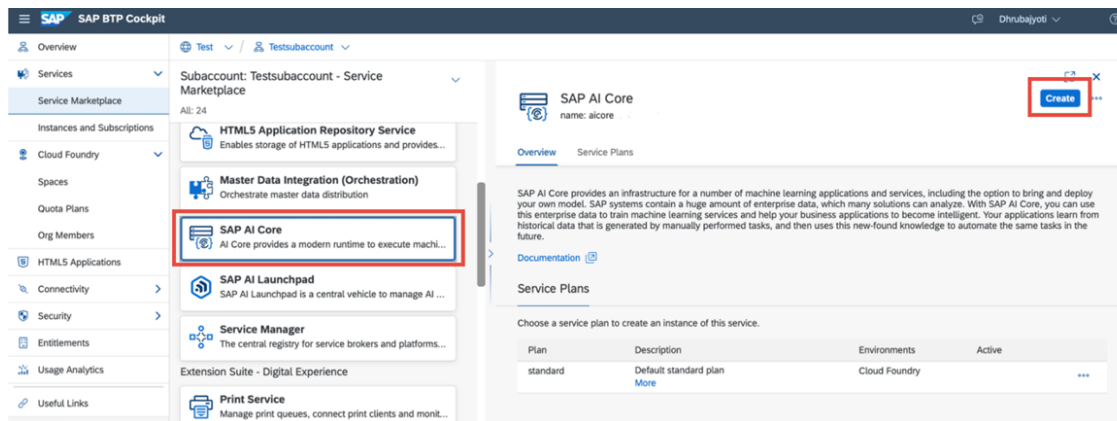
### Procedure

- In SAP BTP cockpit, navigate to your subaccount within your global account and choose *Service Marketplace*.

You will see a tile for SAP AI Core.



- Open the tile and click *Create*.



3. Enter a name for your service instance and choose *Next* (all other details will be filled by default).

### New Instance or Subscription

1 **Basic Info** 2 **Parameters** 3 **Review**

Enter basic info for your instance or subscription.

Service: \* ⓘ  
SAP AI Core

Plan: \*  
standard

Runtime Environment: \*  
Cloud Foundry

Space: \*  
TestAIFoundation

Instance Name: \* ⓘ  
TestAICore

[Next >](#) [Create](#) [Cancel](#)

4. At present, the JSON file upload feature is not supported. Choose *Next* to proceed.

## New Instance or Subscription

1 Basic Info      2 Parameters      3 Review

Configure instance parameters. ⓘ

Upload a JSON file:

[Browse...](#)

Or specify the parameters in JSON format: [Clear](#)

1	
---	--

< Back    **Next >**    Create    Cancel

5. Check the data and choose *Create*.

## New Instance or Subscription

1 Basic Info      2 Parameters      3 Review

Review and verify the instance details.

TestAICore

Service: SAP AI Core

Service Plan: standard

Runtime Environment: Cloud Foundry

Space: TestAIFoundation

*Creating an instance might take a while.*

[Back](#) [Create](#) [Cancel](#)

## Results

When your service instance is created, you can view it on the *Instances and Subscriptions* page of your subaccount.

Subaccount: Testsubaccount - Instances and Subscriptions

To manage the Cloud Foundry user-provided service instances, navigate to Cloud Foundry - Spaces, select your space, and then from Services select Service Instances.

Search [ ] All Services [ ] All Plans [ ] All Statuses [ ]

Subscriptions (0) Instances (1) Environments (1)

Instances (1)

Service instances created in: Cloud Foundry | Kyma/Kubernetes | Other environments

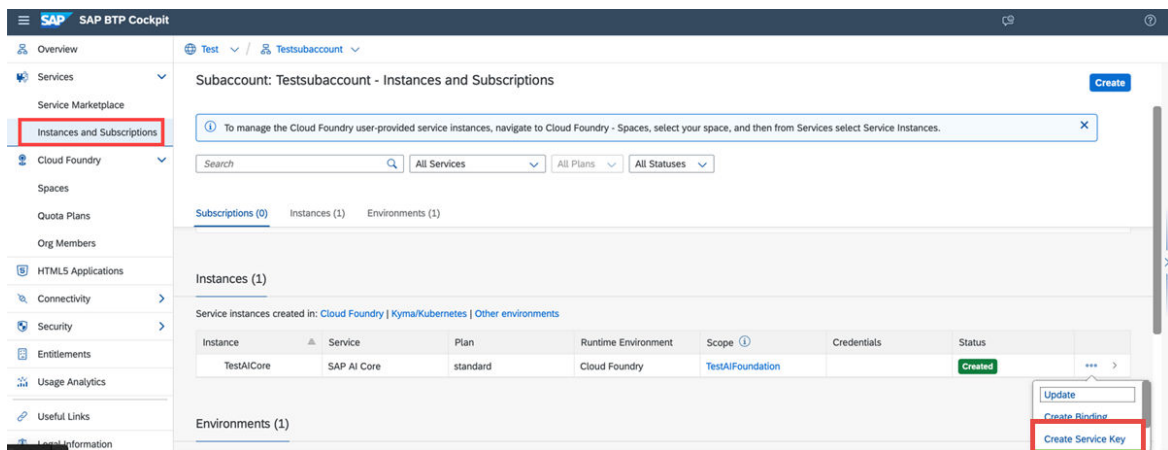
Instance	Service	Plan	Runtime Environment	Scope	Credentials	Status
TestAICore	SAP AI Core	standard	Cloud Foundry	TestAIFoundation		Created

Environments (1)

## 4.1.6 Create a Service Key

### Procedure

1. On the *Instances and Subscriptions* page, find your new instance and choose *Create Service Key* from the dropdown.



The screenshot shows the SAP BTP Cockpit interface. The left sidebar has a red box around 'Instances and Subscriptions'. The main content area shows a table of instances. The first instance is 'TestAICore' with status 'Created'. A dropdown menu is open for this instance, with 'Create Service Key' highlighted in red.

Instance	Service	Plan	Runtime Environment	Scope	Credentials	Status
TestAICore	SAP AI Core	standard	Cloud Foundry	TestAIFoundation		Created

2. Enter a name for your service key and click *Create*.

## New Service Key

Service Key Name: \*

TestAICore

Configure Binding Parameters: ⓘ

Upload a JSON file:

Select JSON file Browse...

Or specify the parameters in JSON format: Clear

1

Create Cancel

### → Remember

You are responsible for the rotation of your access credentials and certificates of SAP AI Core within BTP according to regional policy.

3. **Optional:** To use an x.509 certificate instead of client secret credentials, specify the credentials by adjusting and uploading the following JSON code:

```
{
  "xsuaa": {
    "credential-type": "x509",
    "x509": {
      "key-length": 2048,
      "validity": 7,
      "validity-type": "DAYS"
    }
  }
}
```

```
}
```

- `key-length`: The byte length of the generated private key. Default: 2048.
- `validity`: The validity time unit. DAYS, MONTHS and YEARS are supported. Default: DAYS.
- `validity-type`: The number of time units. Default: 7.

The default combination is a key of length 2048, valid for 7 DAYS.

4. Download your service key to save it.

## Results

You now have your service key, which provides URLs and credentials for accessing the SAP AI Core instance through SAP AI Launchpad, SAP AI Core toolkit, a Third-Party API Platform, or curl.

### → Remember

You are responsible for the rotation of your access credentials and certificates of SAP AI Core within BTP according to regional policy.

### Credentials

▼

Form JSON

```
1 {
2   "clientid": "sb-4376bdb7-          b313",
3   "clientsecret": "xjmmtoqzr27omr7qg8cc0j=",
4   "url": "https://          .authentication.sap.hana.ondemand.com",
5   "identityzone": "ai-          ",
6   "identityzoneid": "1eb727a0-          1",
7   "appname": "4          !b313",
8   "serviceurls": {
9     "AI_API_URL": "https://api.ai.          s.ml.hana.ondemand.com"
10  }
11 }
```

[Copy JSON](#) [Download](#) [Close](#)

If you have generated a client secret, your key will include:

- `clientid`, `clientsecret`, and `url` can be used to generate your authentication token.
- `identityzone` and `identityzoneid` represent your tenant ID.
- `appname` provides the service instance details if service instance isolation is implemented.
- `serviceurls` allow you to interact with SAP AI Core once your authentication token has been generated.

- `AI_API_URL`: Unified AI API to handle ML artifacts (such as training, data, models, and deployments) across multiple hyperscalers.

If you have generated an x.509 certificate, your key will include:

- `certificate`
- `certurl` can be used to generate your authentication token.
- `key` your RSA private key.
- `identityzone` and `identityzoneid` represent your tenant ID.
- `appname` provides the service instance details if service instance isolation is implemented.
- `serviceurls` allow you to interact with SAP AI Core once your authentication token has been generated.
  - `AI_API_URL`: Unified AI API to handle ML artifacts (such as training, data, models, and deployments) across multiple hyperscalers.

## 4.1.7 Use a Service Key

After you have created your service key, it can be used by local clients, apps in other spaces, or entities outside your deployment to access SAP AI Core through one of the available interfaces.

## Using a Third-Party API Platform

### Prerequisites

- You have downloaded and installed the API platform of your choice.
- You have familiarized yourself with the documentation and interface of the platform.

### Procedure

1. Download the JSON collection from [https://api.sap.com/api/AI\\_CORE\\_API/overview](https://api.sap.com/api/AI_CORE_API/overview).
2. Import the JSON file to the API platform.
3. After the import is complete, highlight the collection and select the *Authorization* tab.
4. Navigate to *Configure New Token*, enter the credentials from your service key, and save your changes.

#### Note

- The *Token Name* field is your choice of descriptive identifier.
- The *Access Token URL* is labeled *url* in your service key. Add `/oauth/token` to the end of the URL.
- The *Grant Type* should be **Client Credentials**.

If you see an alert relating to the characters in your credentials, ignore it.

### Note

If you have generated a x.509 certificate instead of client secret credentials, you'll need to use your certificate, key and certUrl to create your token.

```
For example: curl --cert <cert.pem> --key <key.pem> -XPOST <certUrl>/oauth/token
-d 'grant_type=client_credentials&client_id=<client id>'
```

5. Select the *Variables* tab, and set your `baseUrl` from your credentials.  
The `baseUrl` is labeled *AI\_API\_URL* in your service key.
6. Choose *Save*.
7. In the *Authorization* tab, choose *Get New Access Token*. and wait for the authentication process.  
When the authentication process is complete, select *Use Token* to finish. Check that the token is stored in your environment variables. If it is not stored automatically, you can copy and paste it to the *Token* field manually.

## Next Steps

To train and deploy your own AI models, follow the procedure in [Administration \[page 85\]](#).

To use generative AI models provided in the generative AI hub, see [Generative AI Hub](#).

## Using Curl

### Prerequisites

curl is likely to be installed on your operating system by default. To check, open a command prompt and enter `curl -v`. If curl isn't installed, download and install it from <https://curl.se/> .

### Note

On macOS, you may need to install jq so that you can follow the curl commands.

1. Install brew from <https://brew.sh/> .
2. In a Terminal session, run `brew install jq` to install jq in your shell environment.

## Procedure

1. Set up your environment as follows:

### For Linux:

```
# XSUAA details
# URLs should be without trailing slash '/'
```

```
export CLIENTID=<clientid> from service key
export CLIENTSECRET=<clientsecret> from service key
export XSUAA_URL=<url> from service key
export AI_API_URL=<AI_API_URL> from service key
```

#### For Windows PowerShell:

```
$env:CLIENTID = <clientid> from service key
$env:CLIENTSECRET = <clientsecret> from service key
$env:XSUAA_URL = <url> from service key
$env:AI_API_URL = <AI_API_URL> from service key
```

#### Note

The `export` command sets the values of your keys to your environment variable, meaning that they will be retained after you close your terminal session. It is possible to set the environment variables without the `export`, for the current session only.

2. Get the XSUAA OAuth Token using `clientid` and `clientsecret` from the service key to call the APIs.

The XSUAA OAuth token is required for authentication when making AI API calls.

#### For Linux:

```
SECRET=`echo -n "$CLIENTID:$CLIENTSECRET" | base64 -i - `
TOKEN=`curl --request POST \
  --url "$XSUAA_URL/oauth/token" \
  --header "Content-Type: application/x-www-form-urlencoded" \
  --data "grant_type=client_credentials" \
  --data "client_id=$CLIENTID" \
  --data "client_secret=$CLIENTSECRET" `
```

#### For Windows PowerShell:

```
$SECRET = $env:CLIENTID + ":" + $env:CLIENTSECRET
$base64SECRET =
[Convert]::ToBase64String([System.Text.Encoding]::UTF8.GetBytes($SECRET))
$TOKENRESPONSE = Invoke-WebRequest -Method Post "$env:XSUAA_URL/oauth/token"
-Headers @{ "Authorization" = "Basic $base64SECRET"; "Content-Type" =
"application/x-www-form-urlencoded" } -Body "grant_type=client_credentials"
$TOKENJSON = $TOKENRESPONSE.Content | ConvertFrom-Json
$TOKEN = $TOKENJSON.access_token
```

#### Note

The token is valid for a limited time. Once it has expired, create a new token, using the same code snippet.

#### Note

If you have generated a x.509 certificate instead of client secret credentials, you'll need to use your certificate, key and `certUrl` to create your token.

For example: `curl --cert <cert.pem> --key <key.pem> -XPOST <certUrl>/oauth/token -d 'grant_type=client_credentials&client_id=<client id>'`

3. Verify that the token has been fetched properly:

```
echo $TOKEN
```

You should see a long string of alphanumeric characters:

```
eyJhbGciOiJSUzI1NiIsImprdSI6Imh0dHBzOi8vYWktYWxwaGEtdmFsaWRhdGlvbi0yLmF1dGh1bn  
RpY2F0aW9uLnNhcC5oYW5hLm9uZGVtYW5kLmNvbS90b2t1b19rZX1zIiwia2lkIjoizGVmYXVsdC1q  
d3Qta2V5LTMyODMxMjg2NCIsInR5cCI6IkpXVCJ94ZGU5YjAxNmQ0MDk5YjlmM...  
.....  
...ALdfbMsHoYTtF6fNFbf3ZQ
```

## Next Steps

To train and deploy your own AI models, follow the procedure in [Administration \[page 85\]](#).

To use generative AI models provided in the generative AI hub, see [Generative AI Hub](#).

## 4.2 Enabling the Service in the Kyma Environment

Enable SAP AI Core using the standard procedures for the SAP BTP Kyma environment.

### Procedure

1. Create a service instance in the Kyma environment.
2. You can then bind the service instance to your application, or you can create a service key to communicate directly with the service instance. See [Using SAP BTP Services in the Kyma Environment](#).

**Task overview:** [Initial Setup \[page 62\]](#)

### Related Information

[Enabling the Service in Cloud Foundry \[page 62\]](#)

[Using SAP BTP Services in the Kyma Environment](#)

# 5 Tutorials

All available missions for SAP AI Core.

AI Use Case	Tutorial Group/Mission	Description
Getting Started	<a href="#">Use Boosters for Free Plan Use of SAP AI Core and SAP AI Launchpad</a> 	Use a booster to quickly provision the SAP AI Core and SAP AI Launchpad services.
Generative AI	<a href="#">Generative AI with SAP AI Core - Setup</a> 	Set up your SAP Business Technology Platform environment to explore SAP AI Core.
Generative AI	<a href="#">Generative AI with SAP AI Core - Orchestration</a> 	Get started with generative AI workflows with different vendors of Large language models in SAP AI Core and learn the fundamentals of promoting and embedding with generative AI SDK in SAP AI Core.
Generative AI	<a href="#">Generative AI with SAP AI Core - Foundation Models</a> 	Explore foundation large language models that are part of SAP AI Core in a number of use cases.
Generative AI	<a href="#">Generative AI with SAP AI Core - Setup</a> 	Set up your SAP Business Technology Platform environment to explore SAP AI Core.
Generative AI	<a href="#">Generative AI with SAP AI Core - Orchestration</a> 	Get started with generative AI workflows with different vendors of Large language models in SAP AI Core and learn the fundamentals of promoting and embedding with generative AI SDK in SAP AI Core.
Generative AI	<a href="#">Generative AI with SAP AI Core - Foundation Models</a> 	Explore foundation large language models that are part of SAP AI Core in a number of use cases.
Getting Started Predictive AI	<a href="#">Predictive AI with SAP AI Core</a> 	Get started with SAP AI Core, learn the fundamentals, create your first predictive AI workflow, and move your machine learning code to a production cloud.

# 6 Administration

Creating secrets for external programs and tools, that are used with SAP AI Core means that you can connect them without compromising your credentials.

Using SAP AI Core alongside external tools such as GitHub, Docker and Amazon Web Services S3 storage leverages the benefits of version control, containerization, and cloud storage. Your content is made available remotely, if you have a stable internet connection.

You only need to complete the administration steps once. However, you can repeat steps if, for example, you want to add or remove a tool.

## Note

You must have completed the initial setup tasks before configuring your SAP AI Core instance.

## Related Information

[Initial Setup \[page 62\]](#)

## 6.1 Manage Your Git Repository

### 6.1.1 Add a Git Repository

You can use your own git repository to version control your SAP AI Core templates. The GitOps onboarding to SAP AI Core instances involves setting up your git repository and synchronizing your content.

## Using Curl

### Prerequisites

- You've completed the initial setup.
- You have access to a git repository over the Internet.
- You've generated a personal access token for your git repository. For more information, see [Create a Personal Access Token](#) .
- If you want to onboard a git repository hosted on GitLab, make sure that the repository URL contains the `.git` suffix.

- Secrets aren't permitted in your repository. If secrets are used, it isn't possible to synchronize content.
- You've completed the initial setup.

### Note

When you synchronize resources, make sure that there are no naming collisions, especially if you use multiple repositories or applications in one tenant. If you experience difficulties during synchronization, we recommend that you use only one repository or application per tenant.

For example, the following repository URLs are all considered the same repository:

- `https://github.com/user/repo`
- `https://github.com/user/repo/`
- `https://github.com/user/REPO/`

## Context

Git repositories are managed by creating personal access tokens and adding them in SAP AI Core. Personal access tokens are a means of allowing and controlling connections to GitHub repositories without compromising your credentials.

### → Remember

You are responsible for the rotation of your access credentials and certificates of SAP AI Core within BTP according to regional policy.

## Procedure

Submit a POST request to the endpoint `{{apiurl}}/v2/admin/repositories` and include your credentials:

```
curl --location --request POST "$AI_API_URL/v2/admin/repositories" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--data-raw '{
  "url": "https://github.com/john/exemplerepo",
  "username": "john",
  "password": "<GIT_PAT_USER_TOKEN>"
}'
```


You specify your unique git repository details as follows:

- `url`: URL of the git repository

### ⚠ Restriction

Only ASCII alphanumerics, digits, and the characters ".", "-", "\_" and "%" are allowed.

- `username`: (Service) user that's accessing the git repository

- `password`: git personal access token. For more information, see [Create a Personal Access Token](#) .

#### → Tip


To share a repository between two tenants, add the repository in SAP AI Core separately for each tenant and provide the **same** `username` and `password`.

## Next Steps

Create an application to sync your folders. For more information, see [Create an Application \[page 91\]](#).

## Using a Third-Party API Platform

### Prerequisites

- You've completed the initial setup.
- You have access to a git repository over the Internet.
- You've generated a personal access token for your git repository. For more information, see [Create a Personal Access Token](#) .
- If you want to onboard a git repository hosted on GitLab, make sure that the repository URL contains the `.git` suffix.
- Secrets aren't permitted in your repository. If secrets are used, it isn't possible to synchronize content.
- You've completed the initial setup.

#### ⓘ Note

When you synchronize resources, make sure that there are no naming collisions, especially if you use multiple repositories or applications in one tenant. If you experience difficulties during synchronization, we recommend that you use only one repository or application per tenant.

For example, the following repository URLs are all considered the same repository:

- `https://github.com/user/repo`
- `https://github.com/user/repo/`
- `https://github.com/user/REPO/`

## Context

Git repositories are managed by creating personal access tokens and adding them in SAP AI Core. Personal access tokens are a means of allowing and controlling connections to GitHub repositories without compromising your credentials.

### → Remember

You are responsible for the rotation of your access credentials and certificates of SAP AI Core within BTP according to regional policy.

## Procedure

Send a POST request to the endpoint `{{apiurl}}/v2/admin/repositories` and include your credentials in JSON format in the *raw* body:

You specify your unique git repository details as follows:

- `url`: URL of the git repository

### ⚠ Restriction

Only ASCII alphanumeric, digits, and the characters ".", "-", "\_", and "%" are allowed.

- `username`: (Service) user that's accessing the git repository
- `password`: git personal access token. For more information, see [Create a Personal Access Token](#) .

### → Tip

To share a repository between two tenants, add the repository in SAP AI Core separately for each tenant and provide the **same** `username` and `password`.

## Next Steps

Create an application to sync your folders. For more information, see [Create an Application \[page 91\]](#).

## 6.1.2 Edit a Git Repository

### Using Curl

#### Procedure

Submit a PATCH request to the endpoint `{{apiurl}}/v2/admin/repositories` and include your changes:

```
curl --location --request PATCH "$AI_API_URL/v2/admin/repositories" \  
--header "Authorization: Bearer $TOKEN" \  
--header 'Content-Type: application/json' \  

```

```
--data-raw '{
  "url": "https://github.com/john/exemplerepo",
  "username": "john",
  "password": "<GIT_PAT_USER_TOKEN>"
}'
```

You specify your unique git repository details as follows:

- `url`: URL of the git repository

#### ⚠ Restriction

Only ASCII alphanumerics, digits, and the characters ".", "-", "\_" and "%" are allowed.

- `username`: (Service) user that's accessing the git repository
- `password`: git personal access token. For more information, see [Create a Personal Access Token](#) ➦ .

#### → Tip

To share a repository between two tenants, add the repository in SAP AI Core separately for each tenant and provide the **same** `username` and `password`.

## Using a Third-Party API Platform

### Procedure

Send a PATCH request to the endpoint `{{apiurl}}/v2/admin/repositories/{{repositoryName}}` and include your changes in the body.

You specify your unique git repository details as follows:

- `url`: URL of the git repository

#### ⚠ Restriction

Only ASCII alphanumerics, digits, and the characters ".", "-", "\_" and "%" are allowed.

- `username`: (Service) user that's accessing the git repository
- `password`: git personal access token. For more information, see [Create a Personal Access Token](#) ➦ .

#### → Tip

To share a repository between two tenants, add the repository in SAP AI Core separately for each tenant and provide the **same** `username` and `password`.

## 6.1.3 Delete a Git Repository

You remove a Git repository from a connection if its URL is invalid or contains errors, or if the repo is no longer required. Once a Git repository is removed, it can no longer be selected as a source repository for an application.

## Using a Third-Party API Platform

Send a DELETE request to the endpoint `{{apiurl}}/v2/admin/repositories/{{repositoryName}}` and include your repository name.

### Using curl

```
curl --location --request DELETE "{{apiurl}}/v2/admin/repositories/{{repositoryName}}" \
```

## Using Curl

### Context

You remove a Git repository from a connection if its URL is invalid or contains errors, or if the repository is no longer required. Once a Git repository is removed, it can no longer be selected as a source repository for an application.

### Procedure

Run the following code:

```
curl --location --request DELETE "{{apiurl}}/v2/admin/repositories/{{repositoryName}}" \
```

## Using a Third-Party API Platform

### Context

You remove a Git repository from a connection if its URL is invalid or contains errors, or if the repository is no longer required. Once a Git repository is removed, it can no longer be selected as a source repository for an application.

## Procedure

Send a DELETE request to the endpoint `{{apiurl}}/v2/admin/repositories/{{repositoryName}}` and include your repository name.

## 6.2 Manage Applications

### 6.2.1 Create an Application

## Using Curl

### Context

After you add your Git repository, create an application to sync the templates in your repository. The first sync takes some time. You can check the application status to see when it completes. After the initial sync, the system syncs the templates automatically every three minutes. You can also request it manually.

#### Note

Do not create applications that attempt to sync the same source. If two apps have the same `repositoryURL`, `revision`, and `path`, syncing will fail.

## Procedure

Submit a POST request to the endpoint `{{apiurl}}/v2/admin/applications` including details of your application:

```
curl --location --request POST "$AI_API_URL/v2/admin/applications" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--data-raw '{
  "applicationName": "my-app",
  "repositoryUrl": "https://github.com/john/exemplerepo",
  "revision": "HEAD",
  "path": "workflows"
}'
```

- `applicationName`: Specify a name for your application. The name must be between 3 and 64 characters long and match `[A-Za-z0-9\-\_]+`.
- `repositoryUrl`: The URL of a registered git repository. The URL is case-sensitive and must match the URL of a registered git repository.

- `revision`: The revision to target. `<HEAD>` refers to the most recent revision.
- `path`: The path to the target folder that contains the templates to be synced.

Because each application refers to a particular path and revision in the repository, you can create multiple applications for the same `repositoryUrl`.

## Results

After the GitOps setup is completed, the templates in your git repository are automatically synced to SAP AI Core. The synchronization runs approximately every three minutes.

## Next Steps

Check the synchronization status of your application by submitting a GET request to `{{apiurl}}/v2/admin/applications/{{appName}}/status`:

```
curl --location --request GET "$AI_API_URL/v2/admin/applications/{{appName}}/status" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json'
```

As `applicationName`, enter the name of your application that you specified when you created the application.

## Using a Third-Party API Platform

### Context

After you add your Git repository, create an application to sync the templates in your repository. The first sync takes some time. You can check the application status to see when it completes. After the initial sync, the system syncs the templates automatically every three minutes. You can also request it manually.

#### Note

Do not create applications that attempt to sync the same source. If two apps have the same `repositoryURL`, `revision`, and `path`, syncing will fail.

## Procedure

Send a POST request to the endpoint `{{apiurl}}/v2/admin/applications` including details of your application:

- `applicationName`: Specify a name for your application. The name must be between 3 and 64 characters long and match `[A-Za-z0-9\-\_\+]`.
- `repositoryUrl`: The URL of a registered git repository. The URL is case-sensitive and must match the URL of a registered git repository.
- `revision`: The revision to target. `<HEAD>` refers to the most recent revision.
- `path`: The path to the target folder that contains the templates to be synced.

Because each application refers to a particular path and revision in the repository, you can create multiple applications for the same `repositoryUrl`.

## Results

After the GitOps setup is completed, the templates in your git repository are automatically synced to SAP AI Core. The synchronization runs approximately every three minutes.

## Next Steps

Check the synchronization status of your application by sending a GET request to `{{apiurl}}/v2/admin/applications/{{appName}}/status`. As `applicationName`, enter the name of your application that you specified when you created the application.

### Output Code

```
{
  "healthStatus": "Healthy",
  "message": "successfully synced (all tasks run)",
  "reconciledAt": "2021-11-23T10:27:49Z",
  "source": {
    "path": "workflows",
    "repoURL": "https://github.com/username/exemplerepo",
    "revision": "db611bb28be3c853d08867c08b52b8f733b4f7bf",
  },
  "syncFinishedAt": "2021-11-23T10:27:49Z",
  "syncResourceStatus": [
    {
      "kind": "ServingTemplate",
      "message": "servingtemplate.ai.sap.com/text-clf-infer-tutorial
configured",
      "name": "text-clf-infer-tutorial",
      "status": "Synced",
    }
  ],
  "syncedStartedAt": "2021-11-23T10:27:48Z",
  "syncStatus": "Synced",
}
```

## Sync an Application Manually

Applications sync with your GitHub repository automatically at intervals of ~3 minutes. Use the endpoint below to manually request a sync: `{{apiurl}}/admin/applications/{{appName}}/refresh`

### 6.2.2 List Applications

#### Using Curl

##### Procedure

Send a GET request to the endpoint `{{apiurl}}/v2/admin/applications`.

### Using a Third-Party API Platform

##### Procedure

Send a GET request to the endpoint `{{apiurl}}/v2/admin/applications`.

### 6.2.3 Edit an Application

#### Using Curl

##### Procedure

Send a PATCH request to the endpoint `{{apiurl}}/v2/admin/applications/{{appName}}` and include your changes in the body.

## Using a Third-Party API Platform

### Procedure

Send a PATCH request to the endpoint `{{apiurl}}/v2/admin/applications/{{appName}}` and include your changes in the body.

## 6.2.4 Delete an Application

### Using Curl

#### Procedure

Send a DELETE request to the endpoint `{{apiurl}}/v2/admin/applications/{{appName}}`.

## Using a Third-Party API Platform

### Procedure

Send a DELETE request to the endpoint `{{apiurl}}/v2/admin/applications/{{appName}}`.

## 6.3 Manage Resource Groups

A resource group is a unique dedicated namespace or workspace environment, where users can create or add configurations, executions, deployments, and artifacts. They are used for running training jobs or model servers.

Resource groups are used to physically isolate machine learning workloads, and to logically isolate related resources for a usage scenario.

When your tenant is onboarded, a default resource group is automatically created. Default resource groups can't be deleted.

As an administrator, you create, edit, or delete resource groups, based on your service consumers and usage scenarios.

Runtime entities such as executions, deployments, configurations, and artifacts belong to a specific resource group and aren't shared across resource groups. Scenarios, executables, and Docker registry secrets are shared by all resource groups within a tenant.

A resource group is also referred to as an instance.

#### → Remember

Your SAP global account can consist of several accounts. Each account can be associated with a tenant. A tenant can contain multiple resource groups. A tenant always contains a default resource group, as well as the resource groups defined for your usage scenarios.

#### ⓘ Note

The maximum number of resource groups is limited at tenant level to 50. If you reach this limit, you receive an error message. To free up space, delete some resource groups. Alternatively, raise a ticket to increase your quota.

For more information, see [Delete a Resource Group \[page 100\]](#).

For more information, see [Delete a Resource Group](#).

[Create a Resource Group \[page 97\]](#)

[Edit a Resource Group \[page 99\]](#)

[Delete a Resource Group \[page 100\]](#)

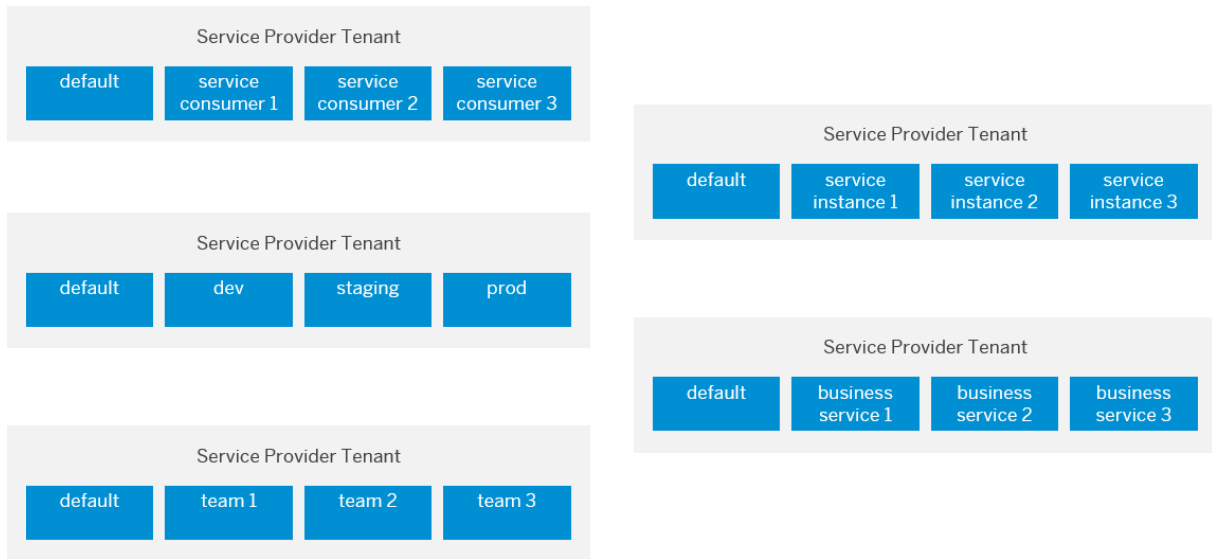
Deletes a resource group that is invalid, contains errors, or is no longer required.

## Resource Group Level Resources

Executables at the tenant level are shared across all resource groups. In contrast, runtime entities such as executions, deployments, configurations, and artifacts belong to a specific resource group and cannot be shared across resource groups. Similarly, generic secrets created within a resource group can be used only for workloads within that group.

You can register an object store at the resource-group level by setting the resource group header. We recommend that you do not use the same object store bucket with the same IAM user for multiple resource groups.

Example resource group mappings are outlined in the figure below:



## 6.3.1 Create a Resource Group

Parent topic: [Manage Resource Groups \[page 95\]](#)

### Related Information

[Edit a Resource Group \[page 99\]](#)

[Delete a Resource Group \[page 100\]](#)

## Using Curl

### Prerequisites

You've completed the initial setup.

You have access to a public-facing Docker registry over the internet. It isn't possible to use a Docker registry behind a VPN or corporate network.

## Context

### ⓘ Note

Resource group Ids must be of length minimum: 3, maximum: 253. The first and last characters must be either a lowercase letter, an uppercase letter, or a number. Character entries from the second to penultimate can include a lower case letter, an upper case letter, a number, a period (.), or a hyphen (-). No other special characters are permitted.

## Procedure

Create a resource group by sending the following:

```
curl --location --request POST "$AI_API_URL/v2/admin/resourceGroups" --header "Authorization: Bearer $TOKEN" --header 'Content-Type: application/json' --data-raw '{ "resourceGroupId": "<ID of your resource group>" }'
```

## Using a Third-Party API Platform

### Prerequisites

You've completed the initial setup.

You have access to a public-facing Docker registry over the internet. It isn't possible to use a Docker registry behind a VPN or corporate network.

## Context

### ⓘ Note

Resource group Ids must be of length minimum: 3, maximum: 253. The first and last characters must be either a lowercase letter, an uppercase letter, or a number. Character entries from the second to penultimate can include a lower case letter, an upper case letter, a number, a period (.), or a hyphen (-). No other special characters are permitted.

## Procedure

1. As the request body, select the *raw* radio button and enter the following:

```
{  
  "resourceGroupId": "<ID of your resource group>"  
}
```

2. Send the request.

## Results

You'll receive a 202 response to confirm that the request to create the resource group has been accepted.

## 6.3.2 Edit a Resource Group

Parent topic: [Manage Resource Groups \[page 95\]](#)

## Related Information

[Create a Resource Group \[page 97\]](#)

[Delete a Resource Group \[page 100\]](#)

## Using Curl

## Context

### Note

Resource group IDs must be of length minimum: 3, maximum: 253. The first and last characters must be either a lowercase letter, an uppercase letter, or a number. Character entries from the second to penultimate can include a lower case letter, an upper case letter, a number, a period (.), or a hyphen (-). No other special characters are permitted.

## Procedure

Create a resource group by sending the following:

```
curl --location --request PATCH "$AI_API_URL/v2/admin/resourceGroups/
{{resource_group_name}}" --header "Authorization: Bearer $TOKEN" --header
'Content-Type: application/json' --data-raw '{ "resourceGroupId": "<ID of your
resource group>"}'
```

## Using a Third-Party API Platform

### Context

#### ⓘ Note

Resource group Ids must be of length minimum: 3, maximum: 253. The first and last characters must be either a lowercase letter, an uppercase letter, or a number. Character entries from the second to penultimate can include a lower case letter, an upper case letter, a number, a period (.), or a hyphen (-). No other special characters are permitted.

## Procedure

Send a PATCH request to the endpoint `{{apiurl}}/v2/admin/resourceGroups/{{resource_group_name}}` with the body:

```
{
  "resourceGroupId": "<ID of your resource group>"
}
```

### 6.3.3 Delete a Resource Group

Deletes a resource group that is invalid, contains errors, or is no longer required.

Parent topic: [Manage Resource Groups \[page 95\]](#)

## Related Information

[Create a Resource Group \[page 97\]](#)

[Edit a Resource Group \[page 99\]](#)

# Using the API

## Context

### → Remember

When you delete a resource group, [Manage mTLS Certificate Secrets \[page 130\]](#) in that group are deleted. Certificates that were issued before the deletion may remain valid until they expire. You are responsible for removing or revoking trust on any external services that rely on those certificates.

## Procedure

Run the following code:

```
curl --location --request POST "$AI_API_URL/v2/admin/resourceGroups/
{{resource_group_name}}"
--header "Authorization: Bearer $TOKEN"
--header 'Content-Type: application/json'
--data-raw '{
"resourceGroupId": "<ID of your resource group>"
}'
```

## Results

Successful responses return code **202** and include a success message.

# Using a Third-Party API Platform

## Context

### → Remember

When you delete a resource group, [Manage mTLS Certificate Secrets \[page 130\]](#) in that group are deleted. Certificates that were issued before the deletion may remain valid until they expire. You are responsible for removing or revoking trust on any external services that rely on those certificates.

## Procedure

Send a DELETE request to the endpoint `{{apiurl}}/v2/admin/resourceGroups/{{resource_group_name}}`.

## 6.4 Manage Object Store Secrets

### 6.4.1 Register an Object Store Secret

Connect SAP AI Core to a cloud object store and manage access using an object store secret. The connected storage stores your dataset, models, and other cache files of the Metaflow Library for SAP AI Core.

Your cloud storage credentials are managed using secrets. Secrets are a means of allowing and controlling connections across directories and tools, without compromising your credentials.

#### ⚠ Restriction

You must create an **object store secret** named **default** to store the training output artifact (for example, a model). If this default object store secret is missing, the training pipeline fails.

For **input training artifacts only**, you can create multiple object store secrets with different names as needed.

#### ⚠ Caution

You are responsible for the rotation of your access credentials and certificates of SAP AI Core within BTP according to regional policy.

To prevent potential data leakage with models themselves, the bucket must be configured to contain only the models and associated data.

You are solely responsible for configuring the object store bucket.

## Using the API

### Context

- Your cloud storage credentials are managed using secrets. Secrets are a means of allowing and controlling connections across directories and tools, without compromising your credentials.
- SAP AI Core supports multiple hyperscaler object stores, including the following:
  - Amazon S3
  - Azure Blob Storage
  - Google Cloud Storage

- OSS (Alibaba Cloud Object Storage Service)
- SAP HANA Cloud, Data Lake

## Procedure

Register your object store secret details using the endpoint `/v2/admin/objectStoreSecrets`.

### Note

For all storage types **except** Azure Blob Storage, all `<data>` fields are required. For Azure, required fields are specified.

- For Amazon S3:

```
curl --location --request POST "$AI_API_URL/v2/admin/objectStoreSecrets" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "S3",
  "bucket": "<S3 bucket name>",
  "endpoint": "<S3 end point>",
  "pathPrefix": "<A path prefix that follows the bucket name>",
  "region": "<S3 region>",
  "data": {
    "AWS_ACCESS_KEY_ID": "<AWS access key ID>",
    "AWS_SECRET_ACCESS_KEY": "<AWS secret access key>"
  }
}'
```

- For OSS (Alibaba Cloud Object Storage Service):

```
curl --location --request POST "$AI_API_URL/v2/admin/objectStoreSecrets" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "oss",
  "pathPrefix": "<path prefix to be appended with bucketname>",
  "data": {
    "BUCKET": "<bucket-name>",
    "ENDPOINT": "oss-cn-shanghai.aliyuncs.com",
    "REGION": "",
    "ACCESS_KEY_ID": "xxxxxx",
    "SECRET_ACCESS_KEY": "xxxxxx"
  }
}'
```

- For SAP HANA Cloud, Data Lake:

```
curl --location --request POST "$AI_API_URL/v2/admin/objectStoreSecrets" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "webhdfs",
  "pathPrefix": "<path prefix to be appended>",

```

```

    "data": {
      // e.g. https://c32727c8-4260-4c37-b97f-edec322dcfa8f.files.hdl.canary-
      eu10.hanacloud.ondemand.com
      "HDFS_NAMENODE": "https://<file-container-name>.files.hdl.canary-
      eu10.hanacloud.ondemand.com",
      "TLS_CERT": "-----BEGIN CERTIFICATE-----
\nMIICmJCCAYIxxxxxxxxxxxxR4wtC32bGO66D+Jc8RhaIA==\n-----END CERTIFICATE-----
\n",
      "TLS_KEY": "-----BEGIN PRIVATE KEY-----
\nMIIEvQIBADANBgkqxxxxxxxxxxxxnor+rtZHhzEfX5dYLC5Pww=\n-----END PRIVATE
KEY-----\n",
      "HEADERS": "{\"x-sap-filecontainer\": \"<file-container-name>\",
\n\"Content-Type\": \"application/octet-stream\"}"
    }
  }
}'

```

### ⚠ Restriction

When using an SAP HANA Data Lake object store with output artifacts pointing to a directory, you can't use `archive: none: {}` in your workflow templates to disable artifact archiving. For more information, see [Workflow Templates](#).

- For Azure Blob Storage:

```

curl --location --request POST "$AI_API_URL/v2/admin/objectStoreSecrets" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "azure",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    "CONTAINER_URI": "https://account_name.blob.core.windows.net/
container_name", //required
    "REGION": "<region name>", //optional
    "CLIENT_ID": "<azure client id>", //optional
    "CLIENT_SECRET": "<azure client secret>", //optional
    "STORAGE_ACCESS_KEY": "sas_token", //required
    "TENANT_ID": "azure tenant id", //optional
    "SUBSCRIPTION_ID": "subscription id", //optional
  }
}'

```

- For Google Cloud Storage (GCS):

```

curl --location --request POST "$AI_API_URL/v2/admin/objectStoreSecrets/
{{objectStoreName}}" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "gcs",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    "BUCKET": "<gcs bucket name>", //
required
    "PRIVATE_KEY": "<base64 encoded service account key>", //
required
  }
}'

```

### → Tip

The `pathPrefix` is useful if you share the same bucket for different projects. You can set the name of your project folder to `my-ml-project1`, for example. All data is then stored in that folder.

### ⓘ Note

If the `AI-Resource-Group` header isn't specified, the `<Resource Group>` is assigned the value `"default"` automatically.

## Results

Successful responses return code **202** and include a success message.

## Using a Third-Party API Platform

### Prerequisites

You've completed the initial setup.

You have access to a public-facing Docker registry over the internet. It isn't possible to use a Docker registry behind a VPN or corporate network.

### Context

- Your cloud storage credentials are managed using secrets. Secrets are a means of allowing and controlling connections across directories and tools, without compromising your credentials.
- SAP AI Core supports multiple hyperscaler object stores, including the following:
  - Amazon S3
  - Azure Blob Storage
  - Google Cloud Storage
  - OSS (Alibaba Cloud Object Storage Service)
  - SAP HANA Cloud, Data Lake

### Procedure

1. Send a POST request to the endpoint `{{apiurl}}/v2/admin/objectStoreSecrets`.
2. As the request body, select the `raw` radio button and enter your object store secret details.

## Note

For all storage types **except** Azure Blob Storage, all `<data>` fields are required. For Azure, required fields are specified.

- For Amazon S3:

```
{
  "name": "default",
  "type": "S3",
  "bucket": "<S3 bucket name>",
  "endpoint": "<S3 end point>",
  "pathPrefix": "<A path prefix that follows the bucket name>",
  "region": "<S3 region>",
  "data": {
    "AWS_ACCESS_KEY_ID": "<AWS access key ID>",
    "AWS_SECRET_ACCESS_KEY": "<AWS secret access key>"
  }
}
```

- For OSS (Alicloud Object Storage Service):

```
{
  "name": "default",
  "type": "oss",
  "pathPrefix": "<path prefix to be appended with bucketname>",
  "data": {
    "BUCKET": "<bucket-name>",
    "ENDPOINT": "oss-cn-shanghai.aliyuncs.com",
    "REGION": "",
    "ACCESS_KEY_ID": "xxxxxx",
    "SECRET_ACCESS_KEY": "xxxxxx"
  }
}
```

- For SAP HANA Cloud, Data Lake:

```
{
  "name": "default",
  "type": "webhdfs",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    // e.g. https://c32727c8-4260-4c37-b97f-ede322dcfa8f.files.hdl.canary-
    eu10.hanacloud.ondemand.com
    "HDFS_NAMENODE": "https://<file-container-name>.files.hdl.canary-
    eu10.hanacloud.ondemand.com",
    "TLS_CERT": "-----BEGIN CERTIFICATE-----
    \nMIICmJCCAYIxxxxxxxxxxxxR4wtC32bGO66D+Jc8RhaIA==\n-----END
    CERTIFICATE-----\n",
    "TLS_KEY": "-----BEGIN PRIVATE KEY-----
    \nMIIEvQIBADANBgkqxxxxxxxxxxxxxxxxnor+rtZHhHzEfX5dYLCs5Pww=\n-----END PRIVATE
    KEY-----\n",
    "HEADERS": "{ \"x-sap-filecontainer\": \"<file-container-name>\",
    \"Content-Type\": \"application/octet-stream\" }"
  }
}
```

## ⚠ Restriction

When using an SAP HANA Data Lake object store with output artifacts pointing to a directory, you can't use `archive: none: {}` in your workflow templates to disable artifact archiving. For more information, see [Workflow Templates](#).

- For Azure Blob Storage:

```
{
  "name": "default",
  "type": "azure",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    "CONTAINER_URI": "https://account_name.blob.core.windows.net/
container_name", //required
    "REGION": "<region name>", //optional
    "CLIENT_ID": "<azure client id>", //optional
    "CLIENT_SECRET": "<azure client secret>", //optional
    "STORAGE_ACCESS_KEY": "sas_token", //required
    "TENANT_ID": "azure tenant id", //optional
    "SUBSCRIPTION_ID": "subscription id", //optional
  }
}
```

- ```
{
  "name": "default",
  "type": "gcs",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    "BUCKET": "<gcs bucket name>", //
    "PRIVATE_KEY": "<base64 encoded service account key>", //
  }
}
```

#### → Tip

The `pathPrefix` is useful if you share the same bucket for different projects. You can set the name of your project folder to `my-ml-project1`, for example. All data is then stored in that folder.

#### ⓘ Note

If the `AI-Resource-Group` header isn't specified, the `<Resource Group>` is assigned the value `"default"` automatically.

3. Send the request.

## 6.4.2 Edit an Object Store Secret

### Using Curl

#### Context

- Your cloud storage credentials are managed using secrets. Secrets are a means of allowing and controlling connections across directories and tools, without compromising your credentials.

- SAP AI Core supports multiple hyperscaler object stores, including the following:
  - Amazon S3
  - Azure Blob Storage
  - Google Cloud Storage
  - OSS (Alibaba Cloud Object Storage Service)
  - SAP HANA Cloud, Data Lake

## Procedure

Edit your object store secret details using the endpoint `$_AI_API_URL/v2/admin/objectStoreSecrets/{objectStoreName}`.

### Note

For all storage types **except** Azure Blob Storage, all `<data>` fields are required. For Azure, required fields are specified.

- For Amazon S3:

```
curl --location --request PATCH "$AI_API_URL/v2/admin/objectStoreSecrets/
{{objectStoreName}}" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "S3",
  "bucket": "<S3 bucket name>",
  "endpoint": "<S3 end point>",
  "pathPrefix": "<A path prefix that follows the bucket name>",
  "region": "<S3 region>",
  "data": {
    "AWS_ACCESS_KEY_ID": "<AWS access key ID>",
    "AWS_SECRET_ACCESS_KEY": "<AWS secret access key>"
  }
}'
```

- For OSS (Alicloud Object Storage Service):

```
curl --location --request PATCH "$AI_API_URL/v2/admin/objectStoreSecrets/
{{objectStoreName}}" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "oss",
  "pathPrefix": "<path prefix to be appended with bucketname>",
  "data": {
    "BUCKET": "<bucket-name>",
    "ENDPOINT": "oss-cn-shanghai.aliyuncs.com",
    "REGION": "",
    "ACCESS_KEY_ID": "xxxxxx",
    "SECRET_ACCESS_KEY": "xxxxxx"
  }
}'
```

- For SAP HANA Cloud, Data Lake:

```
curl --location --request PATCH "$AI_API_URL/v2/admin/objectStoreSecrets/
{{objectStoreName}}" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "webhdfs",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    // e.g. https://c32727c8-4260-4c37-b97f-edec322dcfa8f.files.hdl.canary-
eu10.hanacloud.ondemand.com
    "HDFS_NAMENODE": "https://<file-container-name>.files.hdl.canary-
eu10.hanacloud.ondemand.com",
    "TLS_CERT": "-----BEGIN CERTIFICATE-----
\nMIICmjCCAYIxxxxxxxxxxxxR4wtC32bGO66D+Jc8RhaIA==\n-----END CERTIFICATE-----
\n",
    "TLS_KEY": "-----BEGIN PRIVATE KEY-----
\nMIIEvQIBADANBgkqxxxxxxxxxxxxxxxxnor+rtZHhzhEfX5dYLCS5Pww=\n-----END PRIVATE
KEY-----\n",
    "HEADERS": "{\"x-sap-filecontainer\": \"<file-container-name>\",
\n\"Content-Type\": \"application/octet-stream\"}"
  }
}'
```

- For Azure Blob Storage:

```
curl --location --request PATCH "$AI_API_URL/v2/admin/objectStoreSecrets/
{{objectStoreName}}" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "azure",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    "CONTAINER_URI": "https://account_name.blob.core.windows.net/
container_name", //required
    "REGION": "<region name>", //optional
    "CLIENT_ID": "<azure client id>", //optional
    "CLIENT_SECRET": "<azure client secret>", //optional
    "STORAGE_ACCESS_KEY": "sas_token", //required
    "TENANT_ID": "azure tenant id", //optional
    "SUBSCRIPTION_ID": "subscription id", //optional
  }
}'
```

- For Google Cloud Storage (GCS):

```
curl --location --request PATCH "$AI_API_URL/v2/admin/objectStoreSecrets/
{{objectStoreName}}" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: <Resource group>' \
--data-raw '{
  "name": "default",
  "type": "gcs",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    "BUCKET": "<gcs bucket name>", //
required
    "PRIVATE_KEY": "<base64 encoded service account key>", //
required
  }
}'
```

```
}'
```

### → Tip

The `pathPrefix` is useful if you share the same bucket for different projects. You can set the name of your project folder to `my-ml-project1`, for example. All data is then stored in that folder.

### ⓘ Note

If the `AI-Resource-Group` header isn't specified, the `<Resource Group>` is assigned the value `"default"` automatically.

## Using a Third-Party API Platform

### Context

Your cloud storage credentials are managed using secrets. Secrets are a means of allowing and controlling connections across directories and tools, without compromising your credentials.

SAP AI Core supports multiple hyperscaler object stores, including the following:

- Amazon S3
- Azure Blob Storage
- Google Cloud Storage
- OSS (Alibaba Cloud Object Storage Service)
- SAP HANA Cloud, Data Lake

### Procedure

1. Send a PATCH request to the endpoint .
2. As the request body, select the `raw` radio button and enter your object store secret details.

### ⓘ Note

For all storage types **except** Azure Blob Storage, all `<data>` fields are required. For Azure, required fields are specified.

- For Amazon S3:

```
{
  "name": "default",
  "type": "S3",
  "bucket": "<S3 bucket name>",
  "endpoint": "<S3 end point>",
  "pathPrefix": "<A path prefix that follows the bucket name>",
  "region": "<S3 region>",
  "data": {
```

```

    "AWS_ACCESS_KEY_ID": "<AWS access key ID>",
    "AWS_SECRET_ACCESS_KEY": "<AWS secret access key>"
  }
}

```

- For OSS (Alicloud Object Storage Service):

```

{
  "name": "default",
  "type": "oss",
  "pathPrefix": "<path prefix to be appended with bucketname>",
  "data": {
    "BUCKET": "<bucket-name>",
    "ENDPOINT": "oss-cn-shanghai.aliyuncs.com",
    "REGION": "",
    "ACCESS_KEY_ID": "xxxxxx",
    "SECRET_ACCESS_KEY": "xxxxxx"
  }
}

```

- For SAP HANA Cloud, Data Lake:

```

{
  "name": "default",
  "type": "webhdfs",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    // e.g. https://c32727c8-4260-4c37-b97f-edc322dcfa8f.files.hdl.canary-
    eu10.hanacloud.ondemand.com
    "HDFS_NAMENODE": "https://<file-container-name>.files.hdl.canary-
    eu10.hanacloud.ondemand.com",
    "TLS_CERT": "-----BEGIN CERTIFICATE-----
    \nMIICmJCCAYIxxxxxxxxxxxxR4wtC32bGO66D+Jc8RhaIA==\n-----END
    CERTIFICATE-----\n",
    "TLS_KEY": "-----BEGIN PRIVATE KEY-----
    \nMIIEvQIBADANBgkqxxxxxxxxxxxxxxxxnor+rtZHhzhEfX5dYLCs5Pww=\n-----END PRIVATE
    KEY-----\n",
    "HEADERS": "{ \"x-sap-filecontainer\": \"<file-container-name>\",
    \"Content-Type\": \"application/octet-stream\"}"
  }
}

```

- For Azure Blob Storage:

```

{
  "name": "default",
  "type": "azure",
  "pathPrefix": "<path prefix to be appended>",
  "data": {
    "CONTAINER_URI": "https://account_name.blob.core.windows.net/
    container_name", //required
    "REGION": "<region name>", //optional
    "CLIENT_ID": "<azure client id>", //optional
    "CLIENT_SECRET": "<azure client secret>", //optional
    "STORAGE_ACCESS_KEY": "sas_token", //required
    "TENANT_ID": "azure tenant id", //optional
    "SUBSCRIPTION_ID": "subscription id", //optional
  }
}

```

- {
 

```

        "name": "default",
        "type": "gcs",
        "pathPrefix": "<path prefix to be appended>",
        "data": {

```

```
required      "BUCKET": "<gcs bucket name>", //
required      "PRIVATE_KEY": "<base64 encoded service account key>", //
    }
}
```

#### → Tip

The `pathPrefix` is useful if you share the same bucket for different projects. You can set the name of your project folder to `my-ml-project1`, for example. All data is then stored in that folder.

#### ⓘ Note

If the `AI-Resource-Group` header isn't specified, the `<Resource Group>` is assigned the value "default" automatically.

3. Send the request.

## 6.4.3 Delete an Object Store Secret

### Using Curl

#### Context

Deleting an object store secret stops access to the object store.

#### Procedure

Run the following code:

```
curl --location --request DELETE "$AI_API_URL/v2/admin/objectStoreSecrets/
{{objectStoreName}}" \
```

### Using a Third-Party API Platform

#### Context

Deleting an object store secret stops access to the object store.

## Procedure

Send a DELETE request to the endpoint

## 6.5 Manage Docker Registry Secrets

### 6.5.1 Register Your Docker Registry Secret

Docker packages and runs applications in remote containers. Connect SAP AI Core to a Docker repository and manage access using a Docker registry secret.

## Using Curl

### Prerequisites

You've completed the initial setup.

You have access to a public-facing Docker registry over the internet. It isn't possible to use a Docker registry behind a VPN or corporate network.

### Context

Your Docker credentials are managed using secrets. Secrets allow and control connections across directories and tools without compromising your credentials.

#### → Remember

You are responsible for the rotation of your access credentials and certificates of SAP AI Core within BTP according to regional policy.

Your Docker registry secret lets you authorize SAP AI Core to pull your **private** Docker images from your Docker repository. You specify the name of the secret in your workflows to authenticate the Docker image pull. For more information, see [Workflow Templates](#) and [Serving Templates](#).

## Procedure

1. Submit a POST request to the endpoint `{{apiurl}}/v2/admin/dockerRegistrySecrets`, Include the following parameters in your request body:

- `name`: Set the name of your Docker registry secret. This is your choice of identifier for your secret. In the example, the name is "mydockerregistry".
- `data`: Enter a JSON string that represents your Docker registry secret.

#### Sample Code

```
{
  "name": "mydockerregistry",
  "data": {
    ".dockerconfigjson": "{\"auths\":{\"your.private.registry\":
    {\"username\":\"john\",\"password\":\"docker-accesstoken-or-password\"}}}"
  }
}
```

#### Note

If you are using a public Docker registry from <http://hub.docker.com>, you must provide your Docker URL in the format `https://index.docker.io`, in the `<"auths">` variable input.

#### Sample Code

```
$ curl --location --request POST "$AI_API_URL/v2/admin/
dockerRegistrySecrets" --header "Authorization: Bearer $TOKEN" --header
'Content-Type: application/json' --data-raw '{
  "name": "mydockerregistry",
  "data": {
    ".dockerconfigjson": "{\"auths\": {\"my.docker.repositories.io\":
    {\"username\":\"$USERNAME\", \"password\": \"$PWD\"}}}"
  }
}'
{
  "message": "secret has been created"
}
```

2. After your Docker registry secret has been created, reference it in your template as an image pull secret. For example

#### Source Code

```
spec:
  imagePullSecrets:
  - name: <Name of your Docker registry secret>
```

## Using a Third-Party API Platform

### Prerequisites

You've completed the initial setup.

You have access to a public-facing Docker registry over the internet. It isn't possible to use a Docker registry behind a VPN or corporate network.

## Context

Your Docker credentials are managed using secrets. Secrets allow and control connections across directories and tools without compromising your credentials.

Your Docker registry secret lets you authorize SAP AI Core to pull your **private** Docker images from your Docker repository. You specify the name of the secret in your workflows to authenticate the Docker image pull. For more information, see [Workflow Templates](#) and [Serving Templates](#).

## Procedure

1. Send a POST request to the endpoint `{{apiurl}}/v2/admin/dockerRegistrySecrets`
2. As the request body, select the *raw* radio button and enter the following:

```
{
  "name": "mydockerregistry",
  "data": {
    ".dockerconfigjson": "{\"auths\":{\"your.private.registry\":
    {\"username\":\"john\", \"password\":\"docker-accesstoken-or-password\"}}}"
  }
}
```

- name: Set the name of your Docker registry secret. This is your choice of identifier for your secret. In the example, the name is "mydockerregistry".
  - data: Enter a JSON string that represents your Docker registry secret.
3. Send the request:

### Output Code

```
{
  "message": "secret has been created"
}
```

4. After your Docker registry secret has been created, reference it in your template as an image pull secret.

```
spec:
  imagePullSecrets:
  - name: <Name of your Docker registry secret>
```

## 6.5.2 Edit a Docker Registry Secret

Docker packages and runs applications in remote containers. Connect SAP AI Core to a Docker repository and manage access using a Docker registry secret.

### Using Curl

#### Context

Your Docker credentials are managed using secrets. Secrets allow and control connections across directories and tools without compromising your credentials.

Your Docker registry secret lets you authorize SAP AI Core to pull your **private** Docker images from your Docker repository. You specify the name of the secret in your workflows to authenticate the Docker image pull. For more information, see [Workflow Templates](#) and [Serving Templates](#).

#### Procedure

Submit a PATCH request to the endpoint `$AI_API_URL/v2/admin/dockerRegistrySecrets/{dockerRegistryName}`, Include the following parameters in your request body:

- `name`: Set the name of your Docker registry secret.
- `data`: Enter a JSON string that represents your Docker registry secret.

#### Sample Code

```
{
  "name": "mydockerregistry",
  "data": {
    ".dockerconfigjson": "{\"auths\":{\"your.private.registry\":
    {\\"username\": \"john\", \"password\": \"docker-accesstoken-or-password\"}}}"
  }
}
```

#### Note

If you are using a public Docker registry from <http://hub.docker.com>, you must provide your Docker URL in the format `https://index.docker.io`, in the `<"auths">` variable input.

```
$ curl --location --request PATCH "$AI_API_URL/v2/admin/dockerRegistrySecrets/
{{dockerRegistryName}}" --header "Authorization: Bearer $TOKEN" --header
'Content-Type: application/json' --data-raw '{
  "name": "mydockerregistry",
  "data": {
    ".dockerconfigjson": "{\"auths\": {\"my.docker.repositories.io\":
    {\\"username\": \"$USERNAME\", \"password\": \"$PWD\"}}}"
  }
}'
```

## Using a Third-Party API Platform

### Procedure

Send a PATCH request to the endpoint `{{apiurl}}/v2/admin/repositories/{{repositoryName}}` and include your changes in the body.

You specify your unique git repository details as follows:

- `url`: URL of the git repository

#### ⚠ Restriction

Only ASCII alphanumeric, digits, and the characters ".", "-", "\_" and "%" are allowed.

- `username`: (Service) user that's accessing the git repository
- `password`: git personal access token. For more information, see [Create a Personal Access Token](#) .

#### → Tip

To share a repository between two tenants, add the repository in SAP AI Core separately for each tenant and provide the **same** `username` and `password`.

## 6.5.3 Delete a Docker Registry Secret

Deleting a docker registry secret removes access to the docker registry.

## Using Curl

### Procedure

Submit a DELETE request to the endpoint `$AI_API_URL/v2/admin/dockerRegistrySecrets/{{dockerRegistryName}}`.

## Using a Third-Party API Platform

### Procedure

Send a DELETE request to the endpoint `{{apiurl}}/v2/admin/dockerRegistrySecrets/{{dockerRegistryName}}`

## 6.6 Manage Generic Secrets

### 6.6.1 Create a Generic Secret

A generic secret authorizes SAP AI Core to use your resource group without exposing your credentials.

## Using the API

### Prerequisites

You've completed the initial setup.

You have access to a public-facing Docker registry over the internet. It isn't possible to use a Docker registry behind a VPN or corporate network.

### Context

Generic secrets store sensitive information when system secrets aren't applicable. They're useful in integration scenarios where SAP AI Core acts as an orchestration layer.

#### → Remember

You are responsible for the rotation of your access credentials and certificates of SAP AI Core within BTP according to regional policy.

SAP AI Core lets you use generic secrets at various levels:

- Main-tenant scope
- Tenant-wide level
- Resource-group level

Generic secrets differ from system secrets, like those for object stores or Docker registries. They store sensitive information for the main tenant, all resource groups, or individual resource groups via an API. You can attach these secrets to containers in executions or deployments as environment variables or volume mounts.

To allow the rotation of tenant-wide secrets for long-running deployments without requiring a restart, the deployment must mount the tenant-wide secret. It must also monitor the mounted secret for changes instead of relying on an in-memory copy. When a tenant-wide secret is updated, the tenant must observe the `resourceGroupSecretReplicationStatus` field in the `Get Secret` endpoint to confirm that the secret has been successfully replicated across the required resource groups. For more information, see [Consume Generic Secrets in Executions or Deployments](#).

Each tenant can have a maximum of five tenant-wide secrets. If you reach this limit, you receive an error message. To free up space, delete tenant-wide secrets as described at [Delete a Generic Secret \[page 128\]](#). Alternatively, submit a ticket to request an increase in your quota.

## → Tip

Generic secrets created at the tenant level automatically propagate to all resource groups. However, if a generic secret with the same name is created at the resource-group level, it replaces the tenant-level secret at the time of creation. The system periodically overrides resource-group level secrets with the corresponding tenant-level secret, but this process can take some time. If a resource-group user creates a secret with the same name as an existing tenant-wide secret, it temporarily overwrites the tenant-wide secret at the resource-group level. This behavior can cause issues, especially for critical operations such as metering.

To prevent unintended overwrites, ensure that the tenant prevents resource-group users from creating arbitrary secrets. You can do so in the following ways:

- Restrict users at resource-group level from accessing the `secrets` endpoint by withholding the JWT token.
- Allow users at resource-group level to create generic secrets by making a request using a different authentication mechanism. The main tenant can then validate and transform these requests before propagating them to the runtime adapter, ensuring that secret names remain consistent and critical secrets aren't unintentionally modified.

## Procedure

Send a POST request and enter the URL `AI_API_URL/v2/admin/secrets`.

- **AI-Tenant-Scope** : `true`. The operation will be performed at the main-tenant level.
- **AI-Resource-Group** : `<resource-group-name>`. The operation will be performed at the resource-group level.
- **AI-Tenant-Scope** : `true` and **AI-Resource-Group** : `*`. The operation will be performed at the tenant-wide level.

```
curl --location --request POST "$AI_API_URL/v2/admin/secrets" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: default' \
--data-raw '{
  "name": "my-generic-secret",
  "data": {
    "some-credential": "bXktc2Vuc2l0aXZlLWRhdGE="
  }
}'
```

## 📌 Note

As a convention the secret name can be written without hyphens to make it simple to consume as a Unix environment variable later.

# Using a Third-Party API Platform

## Prerequisites

You've completed the initial setup.

You have access to a public-facing Docker registry over the internet. It isn't possible to use a Docker registry behind a VPN or corporate network.

## Context

Generic secrets store sensitive information when system secrets aren't applicable. They're useful in integration scenarios where SAP AI Core acts as an orchestration layer.

### → Remember

You are responsible for the rotation of your access credentials and certificates of SAP AI Core within BTP according to regional policy.

SAP AI Core lets you use generic secrets at various levels:

- Main-tenant scope
- Tenant-wide level
- Resource-group level

Generic secrets differ from system secrets, like those for object stores or Docker registries. They store sensitive information for the main tenant, all resource groups, or individual resource groups via an API. You can attach these secrets to containers in executions or deployments as environment variables or volume mounts.

To allow the rotation of tenant-wide secrets for long-running deployments without requiring a restart, the deployment must mount the tenant-wide secret. It must also monitor the mounted secret for changes instead of relying on an in-memory copy. When a tenant-wide secret is updated, the tenant must observe the `resourceGroupSecretReplicationStatus` field in the `Get Secret` endpoint to confirm that the secret has been successfully replicated across the required resource groups. For more information, see [Consume Generic Secrets in Executions or Deployments](#).

Each tenant can have a maximum of five tenant-wide secrets. If you reach this limit, you receive an error message. To free up space, delete tenant-wide secrets as described at [Delete a Generic Secret \[page 128\]](#). Alternatively, submit a ticket to request an increase in your quota.

### → Tip

Generic secrets created at the tenant level automatically propagate to all resource groups. However, if a generic secret with the same name is created at the resource-group level, it replaces the tenant-level secret at the time of creation. The system periodically overrides resource-group level secrets with the corresponding tenant-level secret, but this process can take some time. If a resource-group user creates a secret with the same name as an existing tenant-wide secret, it temporarily overwrites the tenant-wide secret at the resource-group level. This behavior can cause issues, especially for critical operations such as metering.

To prevent unintended overwrites, ensure that the tenant prevents resource-group users from creating arbitrary secrets. You can do so in the following ways:

- Restrict users at resource-group level from accessing the `secrets` endpoint by withholding the JWT token.
- Allow users at resource-group level to create generic secrets by making a request using a different authentication mechanism. The main tenant can then validate and transform these requests before propagating them to the runtime adapter, ensuring that secret names remain consistent and critical secrets aren't unintentionally modified.

## Procedure

1. Send a POST request and enter the URL `{{apiurl}}/v2/admin/secrets`.
2. As the request body, select the `raw` radio button and enter your credentials in JSON format:

```
{
  "name": "my-generic-secret",
  "data": {
    "some-credential": "bXktc2VjcmV0LWNyZWRLbnRpYWw=",
    "other-credentials": "bXktc2VjcmV0LW90aGVyLWNyZWRLbnRpYWw="
  }
}
```

- `name`: Set the name of your generic secret.
- `data`: Enter a JSON string that represents your generic secret.

### Note

As a convention the secret name can be written without hyphens to make it simple to consume as a Unix environment variable later.

3. Specify the scope of the request via the header `AI-Tenant-Scope` and `AI-Resource-Group`:
  - `AI-Tenant-Scope : true`. The operation will be performed at the main-tenant level.
  - `AI-Resource-Group : <resource-group-name>`. The operation will be performed at the resource-group level.
  - `AI-Tenant-Scope : true` and `AI-Resource-Group : *`. The operation will be performed at the tenant-wide level.

In this example, we are using the resource-group level.

| Key               | Value                      |
|-------------------|----------------------------|
| AI-Resource-Group | <Your resource group name> |

4. Send the request.

## Results

### Output Code

```
{
  "message": "secret has been created",
  "name": "my-generic-secret"
}
```

## 6.6.2 Get Generic Secrets

Generic secrets can either be retrieved as a single secret, or you can list all existing secrets.

## Get a Secret Using Curl

### Procedure

Submit a GET request to the endpoint `/v2/admin/secrets/<secret-name>`, and include the scope via the headers:

- **AI-Tenant-Scope** : `true`. The operation will be performed at the main-tenant level.
- **AI-Resource-Group** : `<resource-group-name>`. The operation will be performed at the resource-group level.
- **AI-Tenant-Scope** : `true` and **AI-Resource-Group** : `*`. The operation will be performed at the tenant-wide level.

```
curl --location --request GET "$AI_API_URL/v2/admin/secrets/$SECRET_NAME" \
--header "Authorization: Bearer $TOKEN" \
--header 'AI-Resource-Group: default'
```

## Results

The response contains the name, labels, and the creation timestamp of the requested generic secrets. No sensitive information is revealed in the response.

In the case of a tenant-wide secret, the response additionally includes a list of all resource groups associated with the tenant and the current replication status of the secret to these resource groups.

### Output Code

```
{
  "name": "secret-1",
  "createdAt": "<timestamp>",
```

```

"resourceGroupSecretReplicationStatus":{
  "rg-id-1" : true, # secret was replicated correctly in this namespace
  "rg-id-2" : false, # secret was not replicated or does not exist in
this namespace yet
},
"labels":{
  "<key>": "<value>"
}
}
# Example response for tenant-scoped or resource group level secrets:
{
  "name": "secret-1",
  "createdAt": "<timestamp  ",
  "labels": {
    "<key>": "<value>,"
  }
}
}

```

## Get All Secrets Using Curl

### Procedure

Submit a GET request to the endpoint `/v2/admin/secrets`, and include the scope via the headers:

- **AI-Tenant-Scope** : `true`. The operation will be performed at the main-tenant level.
- **AI-Resource-Group** : `<resource-group-name>`. The operation will be performed at the resource-group level.
- **AI-Tenant-Scope** : `true` and **AI-Resource-Group**: `*`. The operation will be performed at the tenant-wide level.

```

curl --location --request GET "$AI_API_URL/v2/admin/secrets" \
--header "Authorization: Bearer $TOKEN" \
--header 'AI-Resource-Group: default'

```

### Results

The response contains the name, labels, and the creation timestamp of the requested generic secrets. No sensitive information is revealed in the response.

## Get a Secret Using a Third-Party API Platform

### Procedure

Send a GET request to the endpoint `{{apiurl}}/v2/admin/secrets/{{secret_name}}`.

- As the request body, select the *none* radio button.

- b. Specify the scope of the request via the header `AI-Tenant-Scope` or `AI-Resource-Group`:
  - `AI-Tenant-Scope : true`. The operation will be performed at the main-tenant level.
  - `AI-Resource-Group : <resource-group-name>`. The operation will be performed at the resource-group level.
  - `AI-Tenant-Scope : true` and `AI-Resource-Group : *`. The operation will be performed at the tenant-wide level.

## Results

The response contains the name, labels, and the creation timestamp of the requested generic secrets. No sensitive information is revealed in the response.

In the case of a tenant-wide secret, the response additionally includes a list of all resource groups associated with the tenant and the current replication status of the secret to these resource groups.

### Output Code

```
{
  "name": "secret-1",
  "createdAt": "<timestamp>",
  "resourceGroupSecretReplicationStatus":{
    "rg-id-1" : true, # secret was replicated correctly in this namespace
    "rg-id-2" : false, # secret was not replicated or does not exist in
this namespace yet
  },
  "labels":{
    "<key>": "<value>",
  }
}
# Example response for tenant-scoped or resource group level secrets:
{
  "name": "secret-1",
  "createdAt": "<timestamp>",
  "labels": {
    "<key>": "<value>",
  }
}
```

## Get All Secrets Using a Third-Party API Platform

### Procedure

Send a GET request to the endpoint `{{apiurl}}/v2/admin/secrets`.

- a. As the request body, select the *none* radio button.
- b. Specify the scope of the request via the header `AI-Tenant-Scope` or `AI-Resource-Group`:
  - `AI-Tenant-Scope : true`. The operation will be performed at the main-tenant level.
  - `AI-Resource-Group : <resource-group-name>`. The operation will be performed at the resource-group level.

- **AI-Tenant-Scope** : **true** and **AI-Resource-Group**: **\***. The operation will be performed at the tenant-wide level.

## Results

The response contains the name, labels, and the creation timestamp of the requested generic secrets. No sensitive information is revealed in the response.

### 6.6.3 Update a Generic Secret

To update a generic secret, use the PATCH endpoint as shown below. The PATCH operation updates the data and/or labels provided. This can be used for rotating secret credentials.

## Using Curl

### Context

Generic secrets store sensitive information when system secrets aren't applicable. They're useful in integration scenarios where SAP AI Core acts as an orchestration layer.

SAP AI Core lets you use generic secrets at various levels:

- Main-tenant scope
- Tenant-wide level
- Resource-group level

Generic secrets differ from system secrets, like those for object stores or Docker registries. They store sensitive information for the main tenant, all resource groups, or individual resource groups via an API. You can attach these secrets to containers in executions or deployments as environment variables or volume mounts.

To allow the rotation of tenant-wide secrets for long-running deployments without requiring a restart, the deployment must mount the tenant-wide secret. It must also monitor the mounted secret for changes instead of relying on an in-memory copy. When a tenant-wide secret is updated, the tenant must observe the `resourceGroupSecretReplicationStatus` field in the `Get Secret` endpoint to confirm that the secret has been successfully replicated across the required resource groups. For more information, see [Consume Generic Secrets in Executions or Deployments](#).

### Procedure

Submit a PATCH request to the endpoint `/v2/admin/secrets/"$SECRET_NAME"`. Specify the scope via the headers:

Specify the scope of the request via the header `AI-Tenant-Scope` and specify the scope via the or `AI-Resource-Group`:

- **AI-Tenant-Scope** : `true`. The operation will be performed at the main-tenant level.
- **AI-Resource-Group** : `<resource-group-name>`. The operation will be performed at the resource-group level.
- **AI-Tenant-Scope** : `true` and **AI-Resource-Group** : `*`. The operation will be performed at the tenant-wide level.

```
curl --location --request PATCH "$AI_API_URL/v2/admin/secrets/$SECRET_NAME" \
--header "Authorization: Bearer $TOKEN" \
--header 'Content-Type: application/json' \
--header 'AI-Resource-Group: default' \
--data-raw '{
  "data": {
    "some-credential": "bXktc2Vuc2l0aXZlLWRhdGE="
  }
  "labels": [
    { "key": "ext.ai.sap.com/<key1>", "value": "<value1>" },
    { "key": "ext.ai.sap.com/<key2>", "value": "<value2>" }
  ]
}'
```

## Results

### Output Code

```
{
  "message": "The secret has been modified",
  "name": "my-generic-secret"
}
```

The response confirms the successful update of the secret. Label updates are applied immediately without requiring secret recreation. You can verify label updates by retrieving the secret using the GET endpoint.

## Using a Third-Party API Platform

### Context

Generic secrets store sensitive information when system secrets aren't applicable. They're useful in integration scenarios where SAP AI Core acts as an orchestration layer.

SAP AI Core lets you use generic secrets at various levels:

- Main-tenant scope
- Tenant-wide level
- Resource-group level

Generic secrets differ from system secrets, like those for object stores or Docker registries. They store sensitive information for the main tenant, all resource groups, or individual resource groups via an API. You can attach these secrets to containers in executions or deployments as environment variables or volume mounts.

To allow the rotation of tenant-wide secrets for long-running deployments without requiring a restart, the deployment must mount the tenant-wide secret. It must also monitor the mounted secret for changes instead of relying on an in-memory copy. When a tenant-wide secret is updated, the tenant must observe the `resourceGroupSecretReplicationStatus` field in the `Get Secret` endpoint to confirm that the secret has been successfully replicated across the required resource groups. For more information, see [Consume Generic Secrets in Executions or Deployments](#).

## Procedure

1. Send a PATCH request to the endpoint `{{apiurl}}/v2/admin/secrets/{{secretName}}`

As the request body, select the `raw` radio button and enter the following code:

### Source Code

```
{
  "data": {
    "updated-credentials": "bXktc2VjcmV0LW90aGVyLWNyZWRLbnRpYWw="
  }
  "labels": [
    { "key": "ext.ai.sap.com/<key1>", "value": "<value1>" },
    { "key": "ext.ai.sap.com/<key2>", "value": "<value2>" }
  ]
}
```

Specify the scope of the request via the header `AI-Tenant-Scope` and specify the scope via the or `AI-Resource-Group`:

- **AI-Tenant-Scope** : `true`. The operation will be performed at the main-tenant level.
- **AI-Resource-Group** : `<resource-group-name>`. The operation will be performed at the resource-group level.
- **AI-Tenant-Scope** : `true` and **AI-Resource-Group** : `*`. The operation will be performed at the tenant-wide level.

You can update labels alongside or instead of secret data. Only labels with the `ext.ai.sap.com/` prefix can be modified.

### ⚠ Restriction

The following labels cannot be updated via PATCH:

- `ext.ai.sap.com/document-grounding`
- `ext.ai.sap.com/documentRepositoryType`

To remove a label, set its value to an empty string (`""`).

2. Send the request.

## Results

### Output Code

```
{
  "message": "The secret has been modified",
  "name": "my-generic-secret"
}
```

The response confirms the successful update of the secret. Label updates are applied immediately without requiring secret recreation. You can verify label updates by retrieving the secret using the GET endpoint.

## 6.6.4 Delete a Generic Secret

To get a secret name, see [Get Generic Secrets \[page 122\]](#).

## Using Curl

### Procedure

Submit a DELETE request to the endpoint `/v2/admin/secrets/"$SECRET_NAME"`. Specify the scope of the request via the header `AI-Tenant-Scope` and `AI-Resource-Group`:

- **AI-Tenant-Scope** : `true`. The operation will be performed at the main-tenant level.
- **AI-Resource-Group** : `<resource-group-name>`. The operation will be performed at the resource-group level.
- **AI-Tenant-Scope** : `true` and **AI-Resource-Group** : `*`. The operation will be performed at the tenant-wide level.

In this example we use the resource-group scope:

```
curl --location --request DELETE "$AI_API_URL/v2/admin/secrets/$SECRET_NAME" \
--header "Authorization: Bearer $TOKEN" \
--header 'AI-Resource-Group: default'
```

## Using a Third-Party API Platform

### Procedure

1. Send a DELETE request to the endpoint `{{apiurl}}/v2/admin/secrets/{{secretName}}`  
As the request body, select the *none* radio button. Specify the scope of the request via the header `AI-Tenant-Scope` or `AI-Resource-Group`:

- **AI-Tenant-Scope** : **true**. The operation will be performed at the main-tenant level.
  - **AI-Resource-Group** : **<resource-group-name>**. The operation will be performed at the resource-group level.
  - **AI-Tenant-Scope** : **true** and **AI-Resource-Group**: **\***. The operation will be performed at the tenant-wide level.
2. Send the request.

## Results

↗ Output Code

```
200
```

## 6.6.5 Consume Generic Secrets in Executions or Deployments

Generic secrets at resource-group level can be attached to containers in executions or deployments. They can either be mounted as a volume or attached as an environment variable. The following examples illustrate how to consume a generic secret in a container by declaring it in the template. Note that only generic secrets can be attached to containers in this way. System secrets can't be consumed in a template.

### Consume a Generic Secret as an Environment Variable

Generic secrets can be attached to containers using either `envFrom.secretRef` or `env.valueFrom.secretKeyRef`:

- Using `envFrom.secretRef`:

```
spec:
  containers:
  - name: my-kserve-container
    image: centaur
    envFrom:
    - secretRef:
      name: my-generic-secret
```

If your secret contains invalid characters, such as hyphens (-), this method results in error. In this case, map your secret to a valid variable name using `env.valueFrom.secretKeyRef`.

- Using `env.valueFrom.secretKeyRef`:

```
spec:
  containers:
  - name: kserve-container
    image: centaur
```

```
env:
- name: my-generic-secret
  valueFrom:
  secretKeyRef:
    name: my-generic-secret
    key: some-credential
```

## Consume a Generic Secret as a Volume Mount

Generic secrets can also be mounted to containers as volumes:

```
spec:
  containers:
  - name: kserve-container
    image: centaur
    volumeMounts:
  - name: my-generic-secret
    mountPath: "/etc/my-generic-secret"
    readOnly: true
  volumes:
  - name: my-generic-secret
    secret:
      secretName: my-generic-secret
```

## Additional Information

Secret names can be included as parameters in the templates and supplied via AI API configurations:

```
envFrom:
- secretRef:
  name: "{{inputs.parameters.secretName}}"
```

## 6.7 Manage mTLS Certificate Secrets

### Context

mTLS certificate secrets let your SAP AI Core workloads authenticate to external services using mutual TLS (mTLS). SAP AI Core automatically retrieves X.509 client certificates from the SAP BTP Certificate Service on your behalf, eliminating the need for manual certificate management. To establish mTLS connections, you mount this certificate in your executions or deployments.

#### ⚠ Restriction

mTLS certificate secrets are managed via the API and are available through SAP AI Core only.

When you create an mTLS certificate secret, the following procedure occurs:

1. SAP AI Core requests a client certificate from the SAP BTP Certificate Service.
2. The certificate and private key are stored as a TLS secret in your resource group's namespace.
3. The response returns the `subjectDN` (subject Distinguished Name) of the issued certificate.
4. You configure trust for this subject DN on your external target service.
5. Your workload mounts the secret as a volume and uses the certificate files for mTLS authentication.

The certificate's subject DN includes organizational fields set by the BTP Certificate Service (such as `C=`, `O=`, `OU=`) and a `CN=` (common name) field that you can customize.

#### Note

The certificate and private key content are never returned by the API. They're only accessible inside a workload container through a volume mount. The API returns only certificate metadata, such as subject DN, expiry, serial number, and common name.

## Establishing Trust on the Target Service

After creating the secret, you must configure the external target service to trust certificates matching the subject DN returned in the creation response.

#### Note

The exact steps depend on the target service.

To establish trust on the target service, you need to achieve the following:

- Register the SAP Cloud Root CA (or the specific intermediate CA) as a trusted issuer.
- Configure an access policy or trust rule that maps the certificate's subject DN pattern to the appropriate authorization level.

#### Remember

If you delete an mTLS certificate secret or stop using it, you should also remove the trust configuration from the external service. Certificates may remain valid until their original expiry date even after the secret is deleted. For more information, see [Delete an mTLS certificate secret \[page 141\]](#) and [Service Offboarding \[page 187\]](#)

## Certificate Lifecycle and Rotation

mTLS certificates have a limited validity period. You can check the `expiresAt` field via the `GET` endpoint to see when the certificate expires. For more information, see [List mTLS Certificate Secrets \[page 137\]](#).

#### Caution

You're responsible for monitoring certificate expiry and rotating secrets before they expire.

To rotate a certificate, send a `PATCH` request to the secret's endpoint. Rotation generates a new certificate and key with the same subject DN. The old certificate remains valid until its original expiry, ensuring a seamless transition with no downtime. For more information, see [Rotate an mTLS Certificate Secret \[page 139\]](#).

## Recommended Rotation Practices

The following best practices are recommended:

1. Monitor the `expiresAt` field of your mTLS certificate secrets regularly.
2. Rotate the secret well before the expiry date (for example, 30 days in advance).
3. If your workload caches the certificate in memory: restart it after rotation to pick up the new certificate. If your workload reads the certificate from disk on each request: the updated files are propagated automatically.

## Scope and Limitations

- mTLS certificate secrets are scoped to a single resource group. They don't support main-tenant scope, tenant-wide scope, or shared resource groups.
- Each secret produces one certificate. If you need multiple certificates (for example, for different external services), create multiple secrets.
- The `commonName` field is optional and limited to 64 characters. If omitted, the secret name is used.
- Secret names must consist of lowercase alphanumeric characters, hyphens, or dots only, with a maximum length of 252 characters.

## Related Information

[Create an mTLS Certificate Secret \[page 132\]](#)

[List mTLS Certificate Secrets \[page 137\]](#)

[Rotate an mTLS Certificate Secret \[page 139\]](#)

[Consume mTLS Certificate Secrets \[page 135\]](#)

[Get Details of an mTLS Certificate Secret \[page 138\]](#)

[Delete an mTLS certificate secret \[page 141\]](#)

## 6.7.1 Create an mTLS Certificate Secret

Create an mTLS certificate secret so that your workloads can make mutually authenticated TLS (mTLS) requests to external services.

### Prerequisites

Make sure that you completed the initial setup. For more information, see [Initial Setup \[page 62\]](#).

## Context

An mTLS certificate secret instructs SAP AI Core to obtain an X.509 client certificate from the SAP BTP Certificate Service on your behalf. Unlike generic secrets, where you provide the credential payload, SAP AI Core generates both the certificate and the private key.

You can mount the generated certificate in executions or deployments to enable mTLS-authenticated requests to external services that trust the certificate's subject DN.

mTLS certificate secrets are scoped to a single resource group. They do not support main-tenant scope, tenant-wide scope, or shared resource groups.

When you create the secret, the response includes the `subjectDN`. You must configure trust for this subject DN on the external service before authentication succeeds.

You can optionally specify a `commonName`. If you omit it, SAP AI Core uses the secret name as the common name. The common name appears in the Common Name field of the certificate's subject distinguished name.

### Note

Creation is asynchronous. After you send a successful POST request (HTTP 202), the certificate may take a few moments to become available. Use a GET request to verify readiness before you run an execution or deployment that mounts the secret.

## Procedure

1. Send a POST request to the endpoint `AI_API_URL/v2/admin/mtlsCertificateSecrets`.
2. Include the `AI-Resource-Group` header to specify the resource group.

### Minimal request (common name defaults to secret name):

```
curl --location --request POST "$AI_API_URL/v2/admin/mtlsCertificateSecrets" \
--header "Authorization: Bearer $TOKEN" \
--header "Content-Type: application/json" \
--header "AI-Resource-Group: default" \
--data-raw '{
  "name": "my-mtls-cert"
}'
```

### Request with a custom common name:

```
curl --location --request POST "$AI_API_URL/v2/admin/mtlsCertificateSecrets" \
--header "Authorization: Bearer $TOKEN" \
--header "Content-Type: application/json" \
--header "AI-Resource-Group: default" \
--data-raw '{
  "name": "my-mtls-cert",
  "data": {
    "commonName": "my-service-client"
  }
}'
```

### Note

The secret name must match the pattern `^[a-z0-9\-\.\ ]+$` and be at most 252 characters. A common name can be up to 64 characters.

## Results

The service returns HTTP 202 `Accepted` and includes the issued certificate's subject DN.

### Output Code

```
202 Accepted
{
  "name": "my-mtls-cert",
  "message": "...",
  "subjectDN": "C=DE, O=SAP SE, OU=SAP Cloud Platform Clients,
OU=<subaccount-ou>, L=<locality-hash>, CN=my-service-client"
}
```

Use the returned `subjectDN` to configure trust on the external service.

## Using a Third-Party API Platform

### Procedure

1. Send a POST request to `{{apiurl}}/v2/admin/mtlsCertificateSecrets`.
2. Select `raw` for the request body and enter the JSON payload.

```
{
  "name": "my-mtls-cert",
  "data": {
    "commonName": "my-service-client"
  }
}
```

3. Set the `AI-Resource-Group` header to the target resource group name.
4. Send the request.

## 6.7.2 Consume mTLS Certificate Secrets

You can consume mTLS certificate secrets in your workloads to authenticate securely with external services that require mutual TLS.

The mTLS certificate secrets can be mounted as volumes in your execution or deployment containers. The secret provides a certificate and private key in PEM format. Your workload uses these files to send mTLS-authenticated requests to external services.

### Note

mTLS certificate secrets can only be consumed as volume mounts. You can't use them as environment variables.

## Mount an mTLS Certificate Secret as a Volume

To mount the certificate and key files in your workload container, define a volume referencing the secret and a corresponding volume mount in your workflow or serving template.

The secret contains two data keys:

- `tls.crt` — the PEM-encoded X.509 client certificate
- `tls.key` — the PEM-encoded private key

You can project these keys to custom file paths by using the `items` field.

### Example workflow template:

```
apiVersion: argoproj.io/v1alpha1
kind: WorkflowTemplate
metadata:
  name: my-mtls-workflow
  annotations:
    scenarios.ai.sap.com/name: "my-scenario"
    executables.ai.sap.com/name: "my-executable"
spec:
  entrypoint: main
  templates:
    - name: main
      inputs:
        parameters:
          - name: mtlsSecretName
      volumes:
        - name: mtls-cert-volume
          secret:
            secretName: "{{inputs.parameters.mtlsSecretName}}"
            items:
              - key: tls.crt
                path: client.crt
              - key: tls.key
                path: client.key
      container:
        image: my-image:latest
        volumeMounts:
          - name: mtls-cert-volume
            mountPath: /mnt/mtls
            readOnly: true
```

Inside the container, the following files are available:

- `/mnt/mtls/client.crt` — the client certificate
- `/mnt/mtls/client.key` — the private key

Your application can use these files to configure an mTLS-authenticated HTTP client. For example, in Python:

```
import requests
response = requests.get(
    "https://external-service.example.com/api/data",
    cert=("/mnt/mtls/client.crt", "/mnt/mtls/client.key")
)
```

## Parameterize the Secret Name

You can parameterize the secret name so that it's provided during execution creation instead of being hardcoded. Use the `inputs.parameters` mechanism:

```
inputs:
  parameters:
    - name: mtlsSecretName
```

When creating the execution through the AI API, provide the parameter value:

```
{
  "parameterBindings": [
    {
      "key": "mtlsSecretName",
      "value": "my-mtls-cert"
    }
  ]
}
```

## Additional Information

- Only mTLS certificate secrets that belong to the same resource group as the execution or deployment can be mounted. The platform validates this during admission.
- If the referenced secret doesn't exist or doesn't belong to the same resource group, the workload is rejected.
- After certificate rotation, see [Rotate an mTLS Certificate Secret \[page 139\]](#), the mounted files update automatically. If your application caches the certificate in memory, you must restart the workload to use the new certificate.

## 6.7.3 List mTLS Certificate Secrets

Retrieve a list of mTLS certificate secrets in a resource group. The response includes certificate metadata such as the subject DN, expiry date, and serial number. The actual certificate and private key are never returned by the API; they are only accessible by mounting the secret in an execution or deployment.

### Prerequisites

You've created at least one mTLS Certificate Secret. For more information, see [Create an mTLS Certificate Secret \[page 132\]](#).

### Procedure

Submit a GET request to the endpoint `$AI_API_URL/v2/admin/mtlsCertificateSecrets`:

Make sure that you've set the following headers:

| Header            | Value                                                                                         |
|-------------------|-----------------------------------------------------------------------------------------------|
| Authorization     | Bearer \$TOKEN                                                                                |
| AI-Resource-Group | The resource group used in the activation steps                                               |
| \$AI_API_URL      | The base URL of your SAP AI Core environment. You can set the URL as an environment variable. |

#### ↔ Sample Code

```
curl --location --request GET "$AI_API_URL/v2/admin/mtlsCertificateSecrets" \  
--header "Authorization: Bearer $TOKEN" \  
--header 'AI-Resource-Group: default'
```

### Results

The response contains a list of mTLS certificate secret metadata objects.

#### ↔ Output Code

```
200 OK  
{  
  "count": 2,  
  "resources": [  
    {  
      "name": "my-mtls-cert",  
      "createdAt": "2026-03-01T12:00:00Z",  
      "subjectDN": "...",
```

```

    "locality": "<locality-hash>",
    "commonName": "my-service-client",
    "expiresAt": "2024-10-01T12:00:00Z",
    "serialNumber": "aabbccdd..."
  },
  {
    "name": "another-cert",
    "createdAt": "2026-03-02T12:00:00Z",
    "subjectDN": "...",
    "locality": "<locality-hash>",
    "commonName": "another-cert",
    "expiresAt": "2024-10-02T12:00:00Z",
    "serialNumber": "eeff0011..."
  }
]
}

```

## 6.7.4 Get Details of an mTLS Certificate Secret

Retrieve metadata for a single mTLS certificate secret. The response includes certificate metadata such as the subject DN, expiry date, and serial number. The actual certificate and private key are never returned by the API; they are only accessible by mounting the secret in an execution or deployment.

### Prerequisites

You've created at least one mTLS Certificate Secret. For more information, see [Create an mTLS Certificate Secret \[page 132\]](#).

### Procedure

Submit a GET request to the endpoint `AI_API_URL/v2/admin/mtlsCertificateSecrets`:

Make sure that you've set the following headers:

| Header            | Value                                                                                         |
|-------------------|-----------------------------------------------------------------------------------------------|
| Authorization     | Bearer \$TOKEN                                                                                |
| AI-Resource-Group | The resource group used in the activation steps                                               |
| AI_API_URL        | The base URL of your SAP AI Core environment. You can set the URL as an environment variable. |

#### Sample Code

```

curl --location --request GET "$AI_API_URL/v2/admin/mtlsCertificateSecrets" \
--header "Authorization: Bearer $TOKEN" \
--header 'AI-Resource-Group: default'

```

## Results

The response contains certificate metadata. No sensitive information is revealed.

### Output Code

```
200 OK
{
  "name": "my-mtls-cert",
  "createdAt": "2024-07-01T12:00:00Z",
  "subjectDN": "C=DE, O=SAP SE, OU=SAP Cloud Platform Clients,
OU=<subaccount-ou>, L=<locality-hash>, CN=my-service-client",
  "locality": "<locality-hash>",
  "commonName": "my-service-client",
  "expiresAt": "2024-10-01T12:00:00Z",
  "serialNumber": "aabbccdd..."
}
```

## 6.7.5 Rotate an mTLS Certificate Secret

Rotate an mTLS certificate secret to obtain a new certificate with the same subject DN. The old certificate remains valid until its original expiry date.

### Prerequisites

You've created at least one mTLS Certificate Secret. For more information, see [Create an mTLS Certificate Secret \[page 132\]](#).

### Context

Certificate rotation generates a new certificate and private key for the secret while preserving the subject DN. Because the subject DN doesn't change, you don't need to update the trust configuration on your external target service.

After rotation, the secret is updated with the new certificate and key. Running workloads that have mounted the secret as a volume will see the new certificate files after a short propagation delay. If your workload reads the certificate at startup and keeps it in memory, you must restart the workload to pick up the rotated certificate.

### ⚠ Restriction

mTLS certificate secrets are managed via the API and are available through SAP AI Core only.

### → Tip

You can rotate a secret multiple times. Each rotation produces a new certificate with a unique serial number. The previous certificate isn't revoked and remains valid until its expiry date. This overlap allows you to rotate without downtime.

## Procedure

Send a PATCH request to the endpoint `$AI_API_URL/v2/admin/mtlsCertificateSecrets`.

Make sure that you've set the following headers:

| Header            | Value                                                                                         |
|-------------------|-----------------------------------------------------------------------------------------------|
| Authorization     | Bearer \$TOKEN                                                                                |
| AI-Resource-Group | The resource group used in the activation steps                                               |
| \$AI_API_URL      | The base URL of your SAP AI Core environment. You can set the URL as an environment variable. |

### ↔ Sample Code

```
curl --location --request PATCH "$AI_API_URL/v2/admin/mtlsCertificateSecrets" \
--header "Authorization: Bearer $TOKEN" \
--header 'AI-Resource-Group: default'
```

## Results

### ↔ Output Code

```
200 OK
{
  "name": "my-mtls-cert",
  "message": "...",
  "subjectDN": "C=DE, O=SAP SE, OU=SAP Cloud Platform Clients,
OU=<subaccount-ou>, L=<locality-hash>, CN=my-service-client"
}
```

The `subjectDN` remains unchanged. You can verify the rotation by calling the `GET` endpoint and checking that `serialNumber` has changed and `expiresAt` has been extended. For more information, see [Get Details of an mTLS Certificate Secret \[page 138\]](#).

## 6.7.6 Delete an mTLS certificate secret

Delete an mTLS certificate secret and retrieve the certificate's subject DN so that you can revoke trust on external services.

### Prerequisites

To get the secret name, see [List mTLS Certificate Secrets \[page 137\]](#).

### Context

You've completed the initial setup. For more information, see [Initial Setup \[page 62\]](#).

When you delete an mTLS certificate secret, the response includes the `subjectDN` of the certificate. Use this subject DN to revoke trust on the external service where you previously configured it.

If you do not revoke trust, any unexpired certificate issued for the same subject DN (for example, a certificate from before rotation) may still authenticate successfully.

### Procedure

1. Send a DELETE request to the endpoint `AI_API_URL/v2/admin/mtlsCertificateSecrets/$SECRET_NAME`.
2. Include the `AI-Resource-Group` header to specify the resource group.
3. Run the following request:

```
curl --location --request DELETE "$AI_API_URL/v2/admin/mtlsCertificateSecrets/
$SECRET_NAME" \
--header "Authorization: Bearer $TOKEN" \
--header "AI-Resource-Group: default"
```

### Results

The service returns HTTP **200 OK** and the subject DN of the deleted certificate.

```
200 OK
{
  "name": "my-mtls-cert",
  "message": "...",
  "subjectDN": "C=DE, O=SAP SE, OU=SAP Cloud Platform Clients, OU=<subaccount-ou>, L=<locality-hash>, CN=my-service-client"
}
```

### Note

After deleting the secret, remove the trust configuration from your external service. The certificate may remain valid until its original expiry date even though the secret has been deleted from SAP AI Core.

## Get a Secret Using a Third-Party API Platform

### Procedure

1. Send a GET request to the endpoint `$AI_API_URL/v2/admin/mtlsCertificateSecrets/$SECRET_NAME`.
2. As the request body, select the `*none*` radio button.
3. Set the header ``AI-Resource-Group`` to the name of the resource group.
4. Send the request.

### Results




#### Output Code

```
200 OK
{
  "name": "my-mtls-cert",
  "message": "...",
  "subjectDN": "..."
}
```

# 7 APIs and API Extensions

Explore APIs and API extensions that you can use with SAP AI Core.

## APIs

| Resource        | Description                                                                           | More Information                                                                                                        |
|-----------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| AI Core API     | Explore the entities, methods, and endpoints provided by the API for SAP AI Core.     | <a href="#">AI Core API</a>          |
| AI API          | Explore the entities, methods, and endpoints provided by the runtime-agnostic AI API. | <a href="#">AI API swagger specification</a>                                                                            |
| Prompt Registry | Simplify the lifecycle management of prompt templates of your business AI scenarios.  | <a href="#">Prompt Registry API</a>  |
| Orchestration   | Enhance content generation with additional capabilities for business AI scenarios.    | <a href="#">Orchestration API</a>  |



## API Extensions


| Resource                     | Description                                                     | More Information                                                 |
|------------------------------|-----------------------------------------------------------------|------------------------------------------------------------------|
| Dataset Management Extension | Manage your dataset files.                                      | <a href="#">AI API Dataset Extension swagger specification.</a>  |
| Resource Groups Extension    | Access usage information for a tenant using the AI API.         | <a href="#">AI API Admin API swagger specification</a>           |
| Analytics Extension          | Add analytics capabilities to the AI API.                       | <a href="#">AI API Analytics Extension swagger specification</a> |
| Metrics Extension            | Query which capabilities of the metrics endpoint are supported. | <a href="#">AI API Metrics API swagger specification</a>         |

## 8 Libraries and SDKs

Explore additional SDKs and libraries that you can use with SAP AI Core.

SDKs Available with SAP AI Core

| Resource                                                                                                                   | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | More Information                                                                                                                                                  |
|----------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>SAP Cloud SDK for AI : JavaScript</p>  | <p>SAP Cloud SDK for AI is the official SDK for SAP AI Core, generative AI hub, and orchestration.</p> <p>You can use SAP Cloud SDK for AI to:</p> <ul style="list-style-type: none"><li>• Integrate chat completion into your business applications with SAP Cloud SDK for AI.</li><li>• Leverage the generative AI hub capabilities of SAP AI Core such as templating, grounding, data masking, content filtering. For more information, see <a href="#">SAP AI Core</a>.</li><li>• Setup your SAP AI Core instance with SAP Cloud SDK for AI.</li></ul> | <ul style="list-style-type: none"><li>• <a href="#">GitHub Repository</a> ↗</li><li>• <a href="#">Documentation</a> ↗</li><li>• <a href="#">NPM</a> ↗</li></ul>   |
| <p>SAP Cloud SDK for AI : Java</p>      | <p>SAP Cloud SDK for AI is the official SDK for SAP AI Core, generative AI hub, and orchestration.</p> <p>You can use SAP Cloud SDK for AI to:</p> <ul style="list-style-type: none"><li>• Integrate chat completion into your business applications with SAP Cloud SDK for AI.</li><li>• Leverage the generative AI hub capabilities of SAP AI Core such as templating, grounding, data masking, content filtering. For more information, see <a href="#">SAP AI Core</a>.</li><li>• Setup your SAP AI Core instance with SAP Cloud SDK for AI.</li></ul> | <ul style="list-style-type: none"><li>• <a href="#">GitHub Repository</a> ↗</li><li>• <a href="#">Documentation</a> ↗</li><li>• <a href="#">Maven</a> ↗</li></ul> |

| Resource                                                                                                               | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | More Information                                                                                                                                                                                 |
|------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>SAP Cloud SDK for AI : Python</p>  | <p>SAP Cloud SDK for AI is the official SDK for SAP AI Core, generative AI hub, and orchestration. The SDK is composed of three python distributions:</p> <p>You can use sap-ai-sdk-base to access the AI API using Python methods and data structures.</p> <p>You can use sap-ai-sdk-core to interact with SAP AI Core for administration and public lifecycle management.</p> <p>You can use sap-ai-sdk-gen to:</p> <ul style="list-style-type: none"> <li>Integrate native SDK libraries and langchain for accessing models on generative AI hub in SAP AI Core.</li> <li>Leverage the orchestration service of generative AI hub with capabilities such as templating, grounding, data masking, and content filtering. For more information, see <a href="#">Generative AI Hub</a>.</li> </ul> | <ul style="list-style-type: none"> <li><a href="#">Pypi Base</a> 🖱️</li> <li><a href="#">Pypi Core</a> 🖱️</li> <li><a href="#">Pypi Gen</a> 🖱️</li> <li><a href="#">Documentation</a></li> </ul> |
| <p>LiteLLM</p>                                                                                                         | <p>The LLMs in the generative AI hub can be accessed from LiteLLM.</p> <p>LiteLLM is an open-source library that supports over 100 LLMs from various providers. It lets you connect LLM and agentic frameworks to the generative AI hub.</p> <div style="border-left: 2px solid orange; padding-left: 10px; margin-top: 10px;"> <p><b>⚠ Caution</b></p> <p>Do not use LiteLLM Versions 1.82.7 and 1.82.8. These versions have vulnerabilities. Versions 1.82.6 and below are safe to install.</p> </div>                                                                                                                                                                                                                                                                                           | <ul style="list-style-type: none"> <li><a href="#">LiteLLM - Getting Started</a> 🖱️</li> <li><a href="#">Agentic code samples</a> 🖱️</li> </ul>                                                  |

# 9 Content Packages

Explore additional Content Packages for use with SAP AI Core.

Content Packages Available with SAP AI Core

| Resource                                                  | Description                                                                                                                                               | More Information                                                            |
|-----------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Content package for large language models for SAP AI Core | The content package for large language models for SAP AI Core simplifies the deployment of large language models with integrated and automated workflows. | <a href="#">Content package for large language models for SAP AI Core</a> ➔ |
| Data robot package                                        | The content package for DataRobot integration for SAP AI Core.                                                                                            | <a href="#">Data robot package for SAP AI Core</a> ➔                        |
| Computer vision package                                   | The content package for image use cases in SAP AI Core adds image classification and feature extraction and is used with the SAP AI SDK Core.             | <a href="#">Computer vision package for SAP AI Core</a> ➔                   |

# 10 Advanced Features

Explore advanced features, within SAP AI Core.

[AI Content as a Service \[page 147\]](#)

With SAP AI Core, you can publish AI content such as workflows, serving templates, or Docker images as a managed service on the SAP BTP *Service Marketplace*. This allows other tenants to consume your content through standard APIs.

## 10.1 AI Content as a Service

With SAP AI Core, you can publish AI content such as workflows, serving templates, or Docker images as a managed service on the SAP BTP *Service Marketplace*. This allows other tenants to consume your content through standard APIs.

A **service provider** is the main SAP AI Core tenant that publishes AI content as a service. For example, the service provider provides a workflow template for other users.

### Note

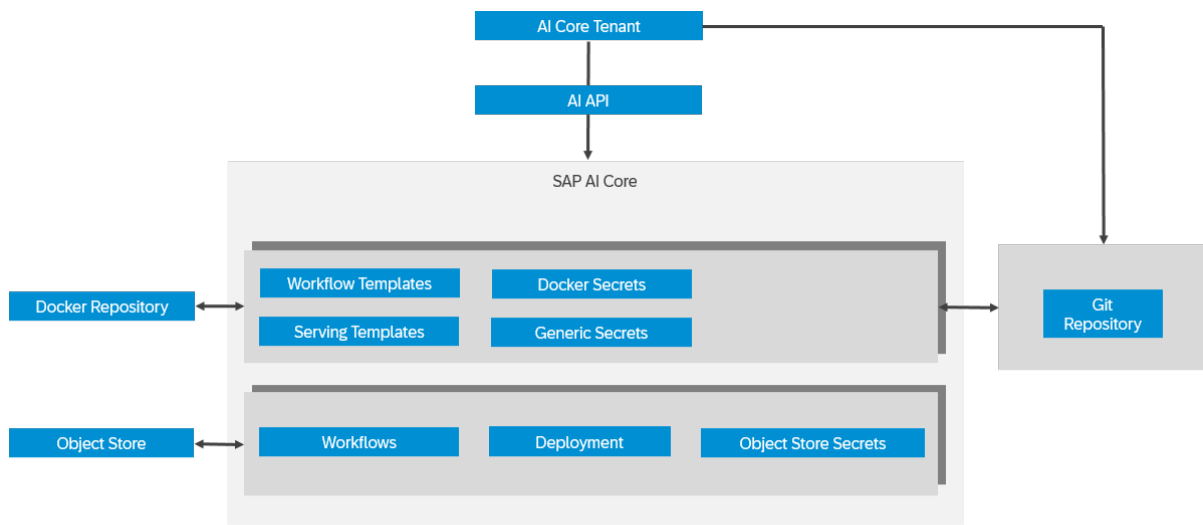
A service provider tenant can provision only one service.

A **service consumer** is the tenant that consumes AI content published by a provider. Consumers create service instances, generate service keys, and use the provided service URL to access the content. They can also start executions or deployments.

## Service Provider Flow

As a service provider, you publish your AI content to the SAP BTP *Service Marketplace* as follows:

1. Create consumer-ready AI content (for example, a workflow, serving template, or Docker image).
2. Create a generic secret for broker registration. For more information, see [Create a Generic Secret \[page 118\]](#).
3. Provide a service custom resource YAML in a registered git repository.
4. Fetch the service broker information by calling the endpoint: `{{apiurl}}/v2/admin/services`.
5. Register the service broker in the SAP BTP *Service Marketplace* and SAP Cloud Management service.

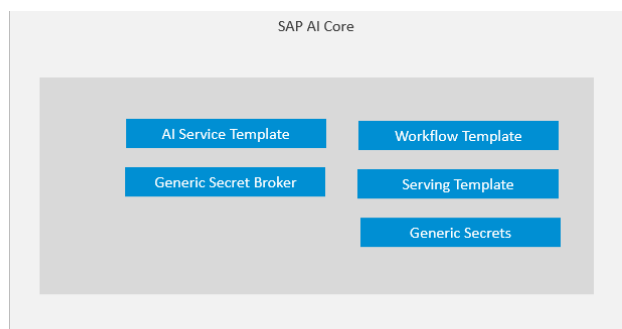
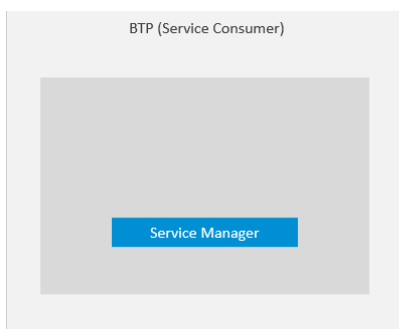


Your AI content is now available in the SAP BTP *Service Marketplace* for consumers to use. The service broker manages onboarding and offboarding of consumers automatically.

## Service Consumer Flow

As a service consumer, you create and use a service instance based on the provider's content.

1. In the SAP BTP *Service Marketplace*, create a service instance. SAP AI Core creates a resource group for you with:  
`<resourceGroupId> == serviceInstanceId.`
2. Create a service key to authenticate against the service.
3. Use the service URL to start executions or deployments with the AI API (`serviceUrl`).



Parent topic: [Advanced Features \[page 147\]](#)

## 10.1.1 Service Custom Resource

The service provider main tenant needs to prepare the service custom resource. The custom resource contains service details, reference to broker credentials or secrets, and capabilities configured for service consumers.

An example service custom resource is provided in the following code block:

```
apiVersion: ai.sap.com/v1alpha1
kind: Service
metadata:
  name: sample-service
spec:
  brokerSecret:
    name: broker-credentials
    usernameKeyRef: username
    passwordKeyRef: password
  description: Service used for demos
  capabilities:
    basic:
      staticDeployments: true
      userDeployments: true
      createExecutions: true
      userPromptTemplates: true
    logs:
      executions: true
      deployments: true
  serviceCatalog:
  - extendCredentials:
      shared:
        serviceUrls:
          AI_API_URL: https://api.ai.internalprod.eu-central-1.aws.ml.hana.ondemand.com
    extendCatalog:
      name: sample-service
      id: sample-service-broker-id
      description: sample service
      bindable: true
      plans:
      - id: sample-service-standard
        description: Standard plan for sample service
        name: standard
        free: false
        metadata:
          supportedPlatforms:
            - cloudfoundry
            - kubernetes
            - sapbtp
```

This can be used as a guide, with values amended as required. To create YAML descriptors, use any text editor with a YAML plugin.

For details about parameters, refer to the following table:

Service Parameter Details

| Type           | Parameter         | Description                                                                                                                                                                         |                                                                                                                                                                                                                                           |
|----------------|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| metadata       | name              | Name for the service                                                                                                                                                                |                                                                                                                                                                                                                                           |
| brokerSecret   | name              | Secret's name containing credentials to register Service Broker.                                                                                                                    |                                                                                                                                                                                                                                           |
|                |                   | <div style="border: 1px solid #ccc; padding: 5px; background-color: #f9f9f9;"> <p><b>Note</b></p> <p>It is mandatory to have this secret registered as a generic secret.</p> </div> |                                                                                                                                                                                                                                           |
|                | usernameKeyRef    | Key reference for username from registered Secret.                                                                                                                                  |                                                                                                                                                                                                                                           |
|                | passwordKeyRef    | Key reference for password from registered Secret.                                                                                                                                  |                                                                                                                                                                                                                                           |
| capabilities   | basic             | staticDeployments                                                                                                                                                                   | Consumers can use existing deployments, but not allowed to Create, Update or Delete (default: true).                                                                                                                                      |
|                |                   | userDeployments                                                                                                                                                                     | Consumers can create deployments (default: true).                                                                                                                                                                                         |
|                |                   | createExecutions                                                                                                                                                                    | Consumers can create executions (default: true).                                                                                                                                                                                          |
|                |                   | userPromptTemplates                                                                                                                                                                 | Consumers can create prompt templates (default: true).                                                                                                                                                                                    |
|                | logs              | executions                                                                                                                                                                          | Consumers can access execution logs.                                                                                                                                                                                                      |
|                |                   | deployments                                                                                                                                                                         | Consumers can access deployment logs.                                                                                                                                                                                                     |
|                |                   | by default                                                                                                                                                                          | Consumer has access to: <ul style="list-style-type: none"> <li>• Create, read artifacts &amp; configurations</li> <li>• Download, upload and delete datasets</li> <li>• Create, read, update &amp; delete object store secrets</li> </ul> |
| serviceCatalog | extendCredentials | Used to mention to be consumed Service URL. It will reflect in the service-key.                                                                                                     |                                                                                                                                                                                                                                           |
|                | extendCatalog     | Used to extend service catalog.                                                                                                                                                     |                                                                                                                                                                                                                                           |

| Type | Parameter                 | Description                                                                                                                       |
|------|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
|      | enableSharedResourceGroup | Flag used to enable or disable share resource group. For more information, see <a href="#">Shared Resource Group [page 154]</a> . |

### ⚠ Restriction

Updating of Service Custom Resource is not Supported

Once a service custom resource has synced successfully, updates of any parameters in templates will not take any effect.

## 10.1.2 Getting Started as a Service Provider

To onboard a service, complete the following:

1. Create a brokerSecret to be used as credentials when registering the service broker.
  1. Use a new POST request to URL `{{apiurl}}/v2/admin/secrets`
  2. Provide credentials (base64 encoded) in the request body in JSON format:

```
{
  "name": "broker-credentials",
  "data": {
    "username": "bXktc2VjcmV0LWNyZWRLbnRpYWw=",
    "password": "bXktc2VjcmV0LW90aGVyLWNyZWRLbnRpYWw="
  }
}
```

3. Specify the scope of the request in the header: `AI-Tenant-Scope: true`
  4. Send the request.
2. Modify the brokerSecret specification section using these details:

```
apiVersion: ai.sap.com/v1alpha1
kind: Service
metadata:
  name: sample-service
spec:
  brokerSecret:
    name: broker-credentials
    usernameKeyRef: username
    passwordKeyRef: password
  description: Service used for demos
  capabilities:
    basic:
      staticDeployments: true
      userDeployments: true
      createExecutions: true
    logs:
      executions: true
      deployments: true
  serviceCatalog:
    - extendCredentials:
      shared:
```

```

        serviceUrls:
          AI_SVC_URL: https://api.ai.internalprod.eu-
central-1.aws.ml.hana.ondemand.com
      extendCatalog:
        name: sample-service
        id: sample-service-broker-id
        description: sample service
        bindable: true
        plans:
          - id: sample-service-standard
            description: Standard plan for sample Service
            name: standard
            free: false
            metadata:
              supportedPlatforms:
                - cloudfoundry
                - kubernetes
                - sapbtp

```

### Note

The username and password are key names from the broker credentials in the previous step.

- Update the `spec.serviceCatalog[].extendCredentials` with the service URL you want to provide to the consumer, which will be part of the service key. Provide catalog details under `spec.serviceCatalog[].extendCatalog`.
- Push your service custom resource to your registered GitHub repository, and wait for the sync to be successful.
- Once it has synced, Check the service details by sending a GET request to URL `{{apiurl}}/v2/admin/services`.

```

{
  "count": 1
  "resources": [
    {
      "name": "sample-service",
      "description": "Service used for demos",
      "status": "PROVISIONED",
      "url": "https://aif-xyzabc.servicebroker.internalprod.eu-
central.aws.ml.hana.ondemand.com"
    }
  ]
}

```

Take note of the service broker URL.

- Register the service broker using `smctl` as subaccount-scoped.
  - Test the registration of the service broker first as subaccount-scoped, before you register it globally. Subaccount-scoped means that your service is automatically visible in the catalog of environments where it's registered.

Once `smctl` is installed, login as shown:

```

# env variables
SERVICE_MANAGER_URL=<sm url e.g. https://service-
manager.cfapps.sap.hana.ondemand.com/>
SVC_SUBACCOUNT_USER=<user-with-servicemanager-
role>SVC_SUBACCOUNT_PWD=<password + 2FA>
SERVICE_BROKER_URL=https://aif-xyzabc.servicebroker.internalprod.eu-
central.aws.ml.hana.ondemand.com
SVC_SUBACCOUNT_SUBDOMAIN=<subaccount e.g. subaccountxyz>
SERVICE_BROKER_USER=<broker username provided in secret>
SERVICE_BROKER_PWD=<broker password provided in secret>

```

```
# smctl login
smctl login -a $SERVICE_MANAGER_URL \
  --param subdomain=$SVC_SUBACCOUNT_SUBDOMAIN \
  -u=$SVC_SUBACCOUNT_USER \
  -p=$SVC_SUBACCOUNT_PWD
```

2. Register your service-broker by providing the broker-name, URL, and credentials.

```
# register service-broker
smctl register-broker sample-service $SERVICE_BROKER_URL -b
$SERVICE_BROKER_USER:$SERVICE_BROKER_PWD
# service-broker registration should complete successfully
```

With the service-broker registered successfully, the service is available in the [Service Marketplace](#).

```
# assuming you are logged in to Cloud Foundry, provided correct subaccount, get
service plan information
cf marketplace -s sample-service
```

Consumers can now create a service instance and service-key. On creation of service instance, SAP AI Core will create a corresponding resource-group with id = instance id, and the service is now ready for use.

## 10.1.3 Metering

Describes how SubaccountID and ServiceInstanceID are available as environment variables in the workflow runtime for metering.

For metering purposes, the workflow runtime automatically injects the following environment variables into each workflow pod:

- `AICORE_CAS_SUBACCOUNT_ID`: The Subaccount ID of the CaaS consumer.
- `AICORE_CAS_SERVICE_INSTANCE_ID`: The Service Instance ID of the CaaS consumer.

These environment variables are available to all processes running inside the workflow pod. You can use them to track usage, or integrate with metering system.

In your workflow container, you can access these variables using the following bash script:

```
echo "Subaccount ID: $AICORE_SUBACCOUNT_ID"
echo "Service Instance ID: $AICORE_SERVICE_INSTANCE_ID"
```

## 10.1.4 Offboarding

To prevent accidental deletion of the service, service providers must provide a deletion strategy as follows:

```
metadata:
  name: sample-service
  annotations:
    ai.sap.com/serviceDeletionStrategy: "delete"
```

With the `serviceDeletionStrategy` annotation, service providers can delete service custom resources from the git repository and proceed for offboarding. For successful service offboarding, all the consumer service instances should be deleted.

## 10.1.5 Shared Resource Group

### Context

When a consumer creates a service instance of the custom resource, a resource group is automatically created with the ID of the service instance ID. A consumer's access to resources is restricted to this resource group.

A shared resource group can be used if a service provider wants to manage a central deployment and make it available to all of the service consumers

#### ⚠ Restriction

Shared resource groups only allow deployment creation for models available in the `foundation-models` scenario, in generative AI hub.

### Procedure

1. To enable shared resource group, set `enableSharedResourceGroup` to `true` in the service custom resource.

#### 📄 Sample Code

```
...
enableSharedResourceGroup: true
serviceCatalog:
...
```

Tenants cannot create, modify or delete a shared resource group using the API. Onboarding for shared resource groups is only possible through the service custom resource..

2. Check the status of your shared resource group by sending a GET request to the endpoint `{{apiurl}}/v2/admin/services/<service-name>`.

#### 📄 Sample Code

```
curl --location '$AI_API_URL/v2/admin/services/aisvc-spam-detection' \
--header 'Authorization: Bearer $TOKEN'
# API Response:
{
  "brokerSecret": {
    "name": "broker-credentials",
    "passwordKeyRef": "password",
```

```

        "usernameKeyRef": "username"
    },
    "capabilities": {
        "basic": {
            "createExecutions": true,
            "staticDeployments": true,
            "userDeployments": true
        },
        "logs": {
            "deployments": true,
            "executions": true
        }
    },
    "description": "Service exposing AI Content for Spam Detection use-
case",
    "name": "aisvc-spam-detection",
    "serviceCatalog": [
        {
            "extendCatalog": {
                "bindable": true,
                "description": "Service exposing AI Content for Spam
Detection use-case",
                "id": "aisvc-spam-detection-0",
                "name": "aisvc-spam-detection",
                "plans": [
                    {
                        "description": "Standard plan for Spam Detection
Service",
                        "free": false,
                        "id": "standard",
                        "metadata": {
                            "supportedPlatforms": [
                                "cloudfoundry",
                                "kubernetes",
                                "sapbtp"
                            ]
                        },
                        "name": "standard"
                    }
                ]
            },
            "extendCredentials": {
                "shared": {
                    "serviceUrls": {
                        "AI_API_URL": "https://api.ai.internalprod.eu-
central-1.aws.ml.hana.ondemand.com"
                    }
                }
            }
        }
    ],
    "sharedResourceGroupStatus": {
        "id": "shared",
        "isEnabled": true,
        "state": "Provisioned"
    },
    "status": "PROVISIONED",
    "statusMessage": "",
    "url": "https://aisvc-425c28c5-aisvc-spam-
detection.servicebroker.internalprod.eu-central-1.aws.ml.hana.ondemand.com"
}

```

## Results

If the `sharedResourceGroupStatus` is `Provisioned`, your shared resource group is provisioned successfully.

## Next Steps

To add resources to your shared resource group, include the header `AI-Resource-Group: shared`. Consumers accessing service must also use the header `AI-Resource-Group: shared` to access the shared resource group.

# 11 Security

Here, we'll explain some of the security aspects of SAP AI Core.

## 11.1 Security Features of Data, Data Flow, and Processes

The table below shows an overview of the data flow for SAP AI Core.

Overview of Data and Security Measures

| Step | Description                                    | Security Measure                                                                                                                                                                                                           |
|------|------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | Transmission control/communication security    | Encrypted (HTTPS) communication. Data-in-transit is encrypted using state-of-the-art TLS settings.                                                                                                                         |
| 2    | Application data residing in persistence layer | Encrypted at rest with state-of-the-art encryption keys generated and maintained by SAP.                                                                                                                                   |
| 3    | Application data residing in persistence layer | Backup and restore capabilities are implemented and tested regularly. Backups are stored in remote locations and backups are encrypted at rest with state-of-the-art encryption keys generated and maintained by SAP.      |
| 4    | Access control and separation by purpose       | Roles and scopes are available for implementing access control. Customer admin can use standard BTP security administration capabilities to assign roles to users to ensure "least privilege" and "segregation of duties". |

## 11.2 Encryption in Transit

Communication with the service, including data upload and download, is encrypted using the transport layer security (TLS) protocol. SAP services support only the latest protocol versions (that is, TLS v1.2 and later) and strong cipher suites. Your systems must use the supported protocol versions and cipher suites to set up secure communication with the services. They must also validate the certificates against the services' domain names to avoid man-in-the-middle attacks.

## 11.3 Authentication and Administration

SAP AI Core uses the authentication mechanisms provided by SAP Authorization and Trust Management service (XSUAA). The credentials to authenticate against the XSUAA to access SAP AI Core are provided as part of the service key for the SAP AI Core service.

For information about SAP Authorization and Trust Management service (XSUAA) in SAP BTP, see [What Is the SAP Authorization and Trust Management Service?](#)

## 11.4 Docker Images

SAP AI Core supports tenant-specific Docker registries (registered via the administration APIs). Additional tenant workloads, such as for execution and deployments, can be created by referencing the Docker images from this Docker registry.

Docker images are cached on virtual machines. These cached Docker images cannot be accessed by other tenants and will not be accessed by SAP.

Cached Docker images are not deleted immediately upon tenant offboarding but are cleaned up as part of operational events such as cluster scaling-down behavior, maintenance, and upgrade of virtual machines.

With every service that you consume, there is a shared security responsibility between you and SAP. Because the creation of a Docker image is the responsibility of the tenant, we strongly recommend that you do not embed or hard-code personal data, sensitive data, or machine learning models inside your Docker images.

For security reasons, Docker containers in SAP AI Core are run as non-root only. For more information, see [Workflow Templates](#).

## 11.5 AI Content Security

AI content covers workflow templates and serving templates, as well as the Docker images used in them. Docker images contain the machine-learning algorithms or code, along with the machine-learning libraries, frameworks, and other dependent packages. Ensure that you follow standard practices for developing secure software when working with AI content.

### Note

Users of AI core are responsible of the content of their docker images and assume the risk of running compromised containers in the platform.

## Security Practices

| Practice                       | Description                                                                                                                                                                          |
|--------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Threat Modeling                | Hold threat modeling workshops to identify and assess security risks or threats in the AI content.                                                                                   |
| Static Code Scans              | Use static code scan (SAST) tools to scan and analyze the code for vulnerabilities.                                                                                                  |
| Open-Source Vulnerability Scan | Scan any open-source components used by the product for vulnerabilities, and patch any vulnerable OSS components found.                                                              |
| Open-Source Strategy           | Develop an update strategy that defines when the open-source components used in a product or service are updated to the latest secure version.                                       |
| Code Reviews                   | Perform peer code reviews of each code change. The reviewer should take a closer look at the code from a security perspective.                                                       |
| Malware Scanning               | Scan for malware in the data uploaded for AI content.                                                                                                                                |
| Secure Code Protection         | Support security assurance starting with source code through deployed service. For example, use Docker image digest and image sign verifications.                                    |
| Docker Base Image Security     | Use a secure, light base image for building the Docker images for the AI content. Ensure you use the latest available base image and remove unused components from the Docker image. |

## 11.6 Kubernetes Security

We recommend that you enable the relevant and applicable Kubernetes security features on your workflow templates and serving templates. Ensure that you enable the appropriate Kubernetes features for your workloads.

### Restricted Pod Settings

To ensure proper security configurations for containers, certain security-related pod fields must adhere to specific values. There is a large number of pod fields, which change over time. Here, we list the most important fields and their required settings. You do not have to explicitly set these values on your own, however if you set these fields incorrectly, the pod will be rejected.

### Setting for `PodSecurityContext`

| Capability | Setting                                     |
|------------|---------------------------------------------|
| Add        | Must not be set                             |
| Drop       | Must contain exactly one element: [ 'all' ] |
| RunAsUser  | Must be set to 65534                        |

| Capability               | Setting                                                  |
|--------------------------|----------------------------------------------------------|
| RunAsGroup               | Must be set to 65534                                     |
| RunAsNonRoot             | Must be set to true. The pod must run as a non-root user |
| AllowPrivilegeEscalation | Must be set to false                                     |

These settings are enforced internally. If your pod is rejected, it is likely due to one or more of these fields being set incorrectly. Please review your configuration and ensure compliance with the above requirements.

## Related Information

[Security Best Practices for Kubernetes Deployment](#) ➔

## 11.7 Configuration Data and Secrets

Workloads can access network resources other than object stores by using credentials at runtime. The ways of including this information at runtime have different standards when it comes to confidentiality.

- For **sensitive** information:  
SAP AI Core allows you to include secrets in the form of generic secrets. Generic secrets are created and managed by the REST APIs in SAP AI Core and consumed securely in a workload.
- For **mTLS authentication** scenarios:  
SAP AI Core provides mTLS certificate secrets. These secrets use managed certificates issued by the SAP BTP Certificate Service. Unlike generic secrets, you do not supply the credential payload, SAP AI Core generates the certificate and private key for you. For more information, see [Manage mTLS Certificate Secrets \[page 130\]](#).
- For **non-sensitive** information:  
You can include non-sensitive parameters using configurations or labels.

### Note

These parameters may be returned in clear text (for example, in GET requests).

## 11.8 Output Encoding

To avoid breaking the business functionality, SAP AI Core does not sanitize any user input. Consumers or applications that consume the AI API are expected to perform necessary output encoding based on the usage context to prevent XSS attacks.

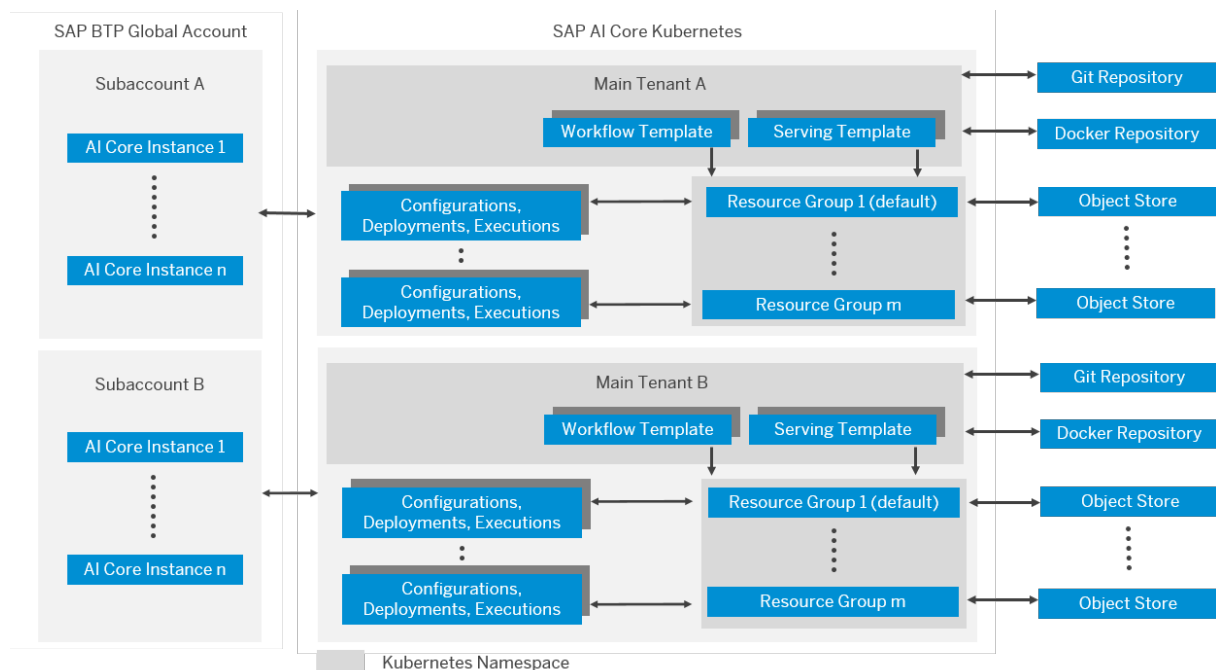
### Related Information

[Cross Site Scripting \(XSS\) – OWASP Site](#)

[Cross Site Scripting Prevention Cheat Sheet – OWASP Site](#)

## 11.9 Multitenancy

SAP AI Core is a tenant-aware BTP reuse service, supporting main tenants and resource groups. Resources are defined for tenants or resource groups as outlined below:



Each main tenant and resource group is mapped to a namespace. The main tenant namespace only contains templates for workflows and model serving. The instances of these objects are created in the respective resource group namespaces and reference the corresponding templates in the main tenant namespace. Each main tenant has a default resource group, which can be used for workloads from the main tenant.

## Tenant-Level Resources

Tenant-level resources include executables such as:

- Workflow templates
- Serving templates
- Docker registries that contain Docker images
- User authentication and authorization (UAA)

User authentication and authorization are based on the SAP AI Core tenant (access token obtained using the service key for SAP AI Core). At runtime or when managing the lifecycle via AI API, the SAP AI Core tenant must set the appropriate resource group in the request header.

## Resource-Group-Level Resources

Executables at the tenant level are shared across all resource groups. In contrast, runtime entities such as executions, deployments, configurations, and artifacts belong to a specific resource group and cannot be shared across resource groups. Similarly, generic secrets created within a resource group can be used only for workloads within that group.

You can register an object store at the resource-group level by setting the resource group header. We recommend that you do not use the same object store bucket with the same IAM user for multiple resource groups.

## Tenant Isolation of Workloads

Workloads run in a sandbox environment and cannot access workflows of other tenants or resource groups. Only TCP is supported for inbound or outbound traffic from a workload. Opening workloads to open Sockets on UDP ports is strongly discouraged. They are not usable, but may pose a theoretical security problem for the workload.

## 11.10 Auditing and Logging Information

Here you can find a list of the security events that are logged by SAP AI Core.

Security Events Written in Audit Logs

| What Events Are Logged          | How to Identify Related Log Events                                                                            |
|---------------------------------|---------------------------------------------------------------------------------------------------------------|
| Creation of object store secret | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded". |

| What Events Are Logged                      | How to Identify Related Log Events                                                                                                                                                   |
|---------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Deletion of object store secret             | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Successful retrieval of object store secret | Successfully read object store connection details for resource group <resource group name> and object store <object store name>                                                      |
| Provisioning of resource group              | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Deprovisioning of resource group            | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Creation of docker registry secret          | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Deletion of docker registry secret          | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Creation of deployments                     | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Deletion of deployments                     | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Creation of executions                      | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Deletion of executions                      | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Creation of ArgoCD application              | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Deletion of ArgoCD application              | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Creation of repositories                    | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Deletion of repositories                    | Message containing a time stamp, tenant IDs, and attributes containing the API call and "new" : " Succeeded " .                                                                      |
| Provisioning of tenant                      | Message containing a time stamp, tenant IDs, and attributes containing "name" : "state" , , "new" : "PENDING" and "name" : "cfAccountState", "new" : "ACTIVE" and success value true |

| What Events Are Logged      | How to Identify Related Log Events                                                                                                                                                                                     |
|-----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Retrieval of tenant         | Message containing <code>Attribute &lt;attribute detail&gt; was read.</code>                                                                                                                                           |
| Deprovisioning of tenant    | Message containing a time stamp, tenant IDs, and attributes containing <code>"name": "state", "new": "PENDING"</code> and <code>"name": "cfAccountState, "new": "DELETED"</code> and success value <code>true</code> . |
| List<br>Get<br>Watch        | Message containing a time stamp, tenant IDs, source IPs, the request URI, and the level.                                                                                                                               |
| Create<br>Update<br>Patch   | Message containing a time stamp, tenant IDs, source IPs, the request URI, the level, and the request and response objects.                                                                                             |
| Delete                      | Message containing a time stamp, tenant IDs, source IPs, the request URI, the level, the and response object.                                                                                                          |
| Token expired               | <code>Jwt is expired</code>                                                                                                                                                                                            |
| No authorization header     | <code>RBAC: access denied</code>                                                                                                                                                                                       |
| Invalid token               | <code>Jwt issuer is not configured</code>                                                                                                                                                                              |
| Wrong <code>tenantId</code> | <code>Jwt verification fails</code>                                                                                                                                                                                    |

The following information is described in the table columns:

- [Event grouping](#) - Events that are logged with a similar format or are related to the same entities.
- [What events are logged](#) - Description of the security or data protection and privacy related event that is logged.
- [How to identify related log events](#) - Search criteria or key words, that are specific for a log event that is created along with the logged event.
- [Additional information](#) - Any related information that can be helpful.

## Related Information

[Audit Logging in the Cloud Foundry Environment](#)

[Audit Logging in the Neo Environment](#)

## 11.11 Data Protection and Privacy

For general information about data protection and privacy on SAP Business Technology Platform, see [Data Protection and Privacy](#).

Data protection is associated with numerous legal requirements and privacy concerns. In addition to compliance with general data protection and privacy acts, it is necessary to consider compliance with industry-specific legislation in different countries/regions. This section describes the specific features and functions that SAP AI Core provides to support compliance with the relevant legal requirements and data privacy.

This guide does not give advice on whether these features and functions are the best method to support company, industry, regional, or country/region-specific requirements. Furthermore, this guide does not give advice or recommendations about additional features that would be required in a particular environment. Decisions related to data protection must be made on a case-by-case basis and under consideration of the given system landscape and the applicable legal requirements.

### Note

SAP does not provide legal advice in any form. SAP software supports data protection compliance by providing security features and specific data protection-relevant functions, such as simplified blocking and deletion of personal data. In many cases, compliance with applicable data protection and privacy laws will not be covered by a product feature. Definitions and other terms used in this document are not taken from a particular legal source.

The extent to which data protection is ensured depends on secure system operation. Network security, security note implementation, adequate logging of system changes, and appropriate usage of the system are the basic technical requirements for compliance with data privacy legislation and other legislation.

For a glossary of Data Protection and Privacy terms in SAP BTP, see the SAP BTP [Glossary for Data Protection and Privacy](#).

### 11.11.1 Data Storage and Processing

SAP AI Core provides functionality that allows you to process data, such as configuration files or Machine Learning (ML) Training or ML Serving.

SAP AI Core acts as the data processor and is not aware of the type of data or category of data. SAP AI Core customers, as Data Controllers, are responsible for fulfilling data protection and privacy (DPP) responsibilities for data storage and processing requirements.

### 11.11.2 Change Logging and Read-Access Logging

SAP AI Core acts as the data processor and is not aware of the type of data or category of data. SAP AI Core customers, as Data Controllers, are responsible for fulfilling data protection and privacy (DPP) responsibilities for data storage and processing requirements.

For any applications or services you develop using SAP AI Core, you must ensure that they include relevant logging functions, and ensure compliance with the data privacy laws by making sure that the data is properly logged.

### 11.11.3 Consent

SAP AI Core acts as the data processor and is not aware of the type of data or category of data. SAP AI Core customers, as Data Controllers, are responsible for fulfilling asking for consent from data subjects before collecting any personal data.

### 11.11.4 Deletion

SAP AI Core supports **Bring your object store**, whereby customers register the object store secret where artifact relevant files (such as training data or machine learning models or other types) are stored. Such data is used by the ML workloads during processing, such as ML Training or ML Serving.

The notion of artifacts is limited to capture the metadata of the data. Data is physically stored in the object store and SAP AI Core is not responsible for deletion of files. When artifacts are deleted using AI API, corresponding metadata are deleted from SAP AI Core service and the actual files are not deleted from the object store. For more information about the AI API, see [AI API Overview \[page 55\]](#).

Upon offboarding, SAP AI Core will clean up the cached data within AI Core used for processing purposes.

SAP AI Core customers, as Data Controllers, are responsible for deletion of data from the registered object store.

### 11.11.5 Security and Customer Data Protection

SAP product standard security and the data protection and privacy (DPP) requirements set high standards and obligations when it comes to securing and protecting customer data that is entrusted to SAP.

Customer data protection is handled in three ways:

- Customer data is imported, output, and processed by the services for no purpose beyond that to which the customer has subscribed.
- Customer data is protected from malicious access by security technologies that include authentication and authorization.
- Customer data is protected from accidental exposure to SAP administrators or support persons by security policies, access controls, and monitoring.

## 12 Accessibility Features in SAP AI Core

To optimize your experience of SAP AI Core, SAP AI Core provides features and settings that help you use the software efficiently.

### Note

SAP AI Core uses SAP AI Launchpad for its interface, which is based on SAPUI5. For this reason, accessibility features for SAPUI5 also apply. See the accessibility documentation for SAPUI5 on SAP Help Portal at [Accessibility for End Users](#).

For more information on-screen reader support and keyboard shortcuts, see [Screen-Reader Support for SAPUI5 Controls](#) and [Keyboard Handling for SAPUI5 Elements](#).

# 13 Monitoring and Troubleshooting

Explore solutions to potential issues, and find out how to get support.

## Getting Support

If you encounter an issue with this service, we recommend that you follow the procedure below.

### Check Platform Status

Check the availability of the platform at [SAP Trust Center](#).

For more information about platform availability, updates and notifications, see [Platform Updates and Notifications in the Cloud Foundry Environment](#).

### Check Guided Answers

In the SAP Support Portal, check the [Guided Answers](#) section for SAP Business Technology Platform. You can find solutions for general SAP Cloud Platform issues as well as for specific services there.

### Contact SAP Support

You can report an incident or error through the [SAP Support Portal](#).

Please use the following component(s) for your incident:

| Component Name | Component Description |
|----------------|-----------------------|
| CA-ML-AIC      | SAP AI Core           |

We recommend that you include the following information when you submit the incident:

- Region information
- Subaccount technical name
- URL of the page where the incident or error occurs
- Steps or clicks used to replicate the error
- Screenshots, videos, or the code entered

## 13.1 Troubleshooting

For troubleshooting information, see the following sections:

[Repository \[page 169\]](#)

[Configuration \[page 170\]](#)

[Artifacts \[page 172\]](#)

[Application \[page 174\]](#)

[Execution \[page 179\]](#)

[Docker \[page 181\]](#)

[Deployment \[page 182\]](#)

[Miscellaneous \[page 183\]](#)

## 13.1.1 Repository

### Repository ra-aicore-test not found for tenant

You get the result:

```
{
  "error": {
    "code": "500",
    "details": [
      {
        "code": null,
        "message": "Repository ra-aicore-test not found for tenant
b82a8318"
      }
    ],
    "message": "Repository ra-aicore-test not found for tenant b82a8318"
  },
  "request_id": null,
  "target": "/api/v4alpha/repositories"
}
```

and:

```
AIAPIServerException: Failed to post /admin/repositories: Repository ra-aicore-
test not found for tenant 68
```

#### Follow the solution:

Use a different name for the value of the name parameter. The exception is raised by reuse of the name aicore-test.

```
response = ai_api_client.rest_client.post(
    path="/admin/repositories",
    body={
        "name": "aicore-test-1",
        "url": "https://github.com/John/aicore-test",
    }
)
print(response)
{'message': 'Repository has been on-boarded.'}
```

Parent topic: [Troubleshooting \[page 168\]](#)

## Related Information

[Configuration \[page 170\]](#)

[Artifacts \[page 172\]](#)

[Application \[page 174\]](#)

[Execution \[page 179\]](#)

[Docker \[page 181\]](#)

[Deployment \[page 182\]](#)

[Miscellaneous \[page 183\]](#)

## 13.1.2 Configuration

### Could not create configuration, because executable `<x>` for scenario `<y>` is not found

When you try to create a configuration, you are told that an executable cannot be found for your scenario.

#### Check the following:

1. Check that you are using the `name` value from your workflow for the executable ID in the configuration.

```
apiVersion: argoproj.io/v1alpha1
kind: WorkflowTemplate
metadata:
  name: text-clf-train-tutorial
  annotations:
    scenarios.ai.sap.com/description: "SAP developers tutorial
scenario"
...
```

#### Note

Do not use the value from `executables.ai.sap.com/id` as an executable ID.

2. Check that you are using the value of `executables.ai.sap.com/id` from your workflows as your scenario ID.

```
...
  artifacts.ai.sap.com/text-data.kind: "dataset"
  artifacts.ai.sap.com/text-model-tutorial.kind: "model"
  labels:
    scenarios.ai.sap.com/id: "text-clf-tutorial"
    ai.sap.com/version: "2.1.0"
  spec:
```

...

## Log message: using minio client

### Check the following:

1. Check that you are using the name value from your workflow for the executable ID in the configuration.

```
apiVersion: argoproj.io/v1alpha1
kind: WorkflowTemplate
metadata:
  name: text-clf-train-tutorial
  annotations:
    scenarios.ai.sap.com/description: "SAP developers tutorial
scenario"
...
```

#### Note

Do not use the value from `executables.ai.sap.com/id` as an executable ID.

2. Check that you are using the value of `executables.ai.sap.com/id` from your workflows as your scenario ID.

```
...
  artifacts.ai.sap.com/text-data.kind: "dataset"
  artifacts.ai.sap.com/text-model-tutorial.kind: "model"
  labels:
    scenarios.ai.sap.com/id: "text-clf-tutorial"
    ai.sap.com/version: "2.1.0"
  spec:
  ...
```

Parent topic: [Troubleshooting \[page 168\]](#)

## Related Information

[Repository \[page 169\]](#)

[Artifacts \[page 172\]](#)

[Application \[page 174\]](#)

[Execution \[page 179\]](#)

[Docker \[page 181\]](#)

[Deployment \[page 182\]](#)

[Miscellaneous \[page 183\]](#)

## 13.1.3 Artifacts

### No output artifact has been generated

#### Complete the following:

Define the `globalName` parameter for the output artifact in your workflow:

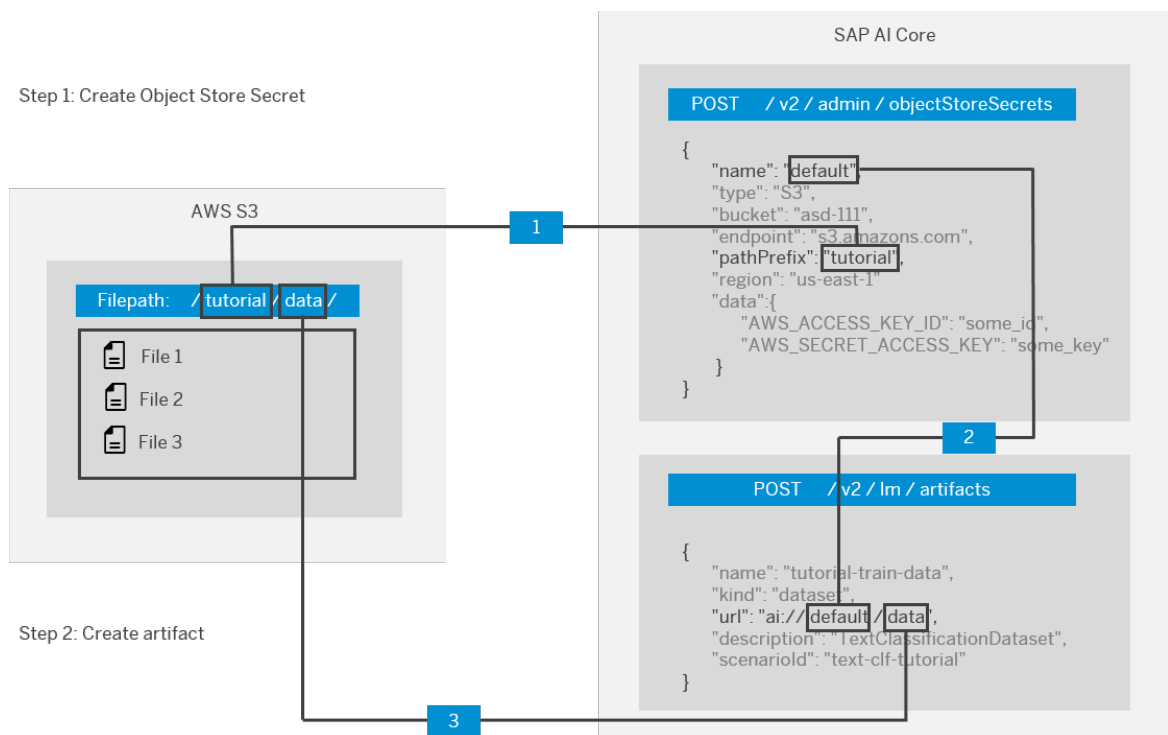
```
...
  executables.ai.sap.com/description: "Text classification Scikit training
executable"
  executables.ai.sap.com/name: "text-clf-train-tutorial-exec"
  artifacts.ai.sap.com/text-data.kind: "dataset"
  artifacts.ai.sap.com/text-model-tutorial.kind: "mind"
labels:
  scenarios.ai.sap.com/id: "text-clf-tutorial"
  ai.sap.com/version: "2.1.0"
...
...
  outputs:
    artifacts:
      -name: text-model-tutorial
      path: /app/model
      global name: text-model-tutorial
      archive:
        none: {}
    container:
...

```

### Failed to load artifacts: The specified key does not exist

#### Complete the following:

1. Ensure you have created an object store secret using the naming convention `<name>` and the `pathPrefix` from your AWS S3 path. Refer to the following diagram:



- When creating the artifact, don't add the trailing forward slash (/) in URL parameter:
  - Incorrect usage: `"url": "ai://yourObjectStoreSecretName/folder/subfolder/"`
  - Correct usage: `"url": "ai://yourObjectStoreSecretName/folder/subfolder"`

Parent topic: [Troubleshooting \[page 168\]](#)

## Related Information

- [Repository \[page 169\]](#)
- [Configuration \[page 170\]](#)
- [Application \[page 174\]](#)
- [Execution \[page 179\]](#)
- [Docker \[page 181\]](#)
- [Deployment \[page 182\]](#)
- [Miscellaneous \[page 183\]](#)

## 13.1.4 Application

### Templates are not synced via applications

After you have deleted or created an application, templates are not synced.

#### Follow the solution:

1. Delete the synced application using the endpoint:  
**DELETE** `{{apiurl}}/v2/admin/applications/{{appName}}`
2. Offboard the associated GitHub repository using the endpoint:  
**DELETE** `{{apiurl}}/v2/admin/repositories/{{repositoryName}}`
3. Onboard the associated GitHub repository using a personal access token instead of your GitHub password.  
For more information, see [Creating a personal access token](#) .  
**POST** `{{apiurl}}/v2/admin/repositories`

#### Sample Code

Body:

```
{
  "name": "aicore-test",
  "url": "https://github.com/john/aicore-test",
  "username": "john",
  "password": "yourGitHubPersonalAccessToken"
}
```

4. Create the application using the endpoint:  
**POST** `{{apiurl}}/v2/admin/applications`
5. Check the ArgoCD applications to determine if the repository has been synchronized correctly for the tenant. For example, check that there are no duplicated workflow names. The value of the name parameter is considered as an executable ID.

```
name: text-clf-train-tutorial
  annotations:
    ...apiVersion: argoproj.io/v1alpha1
kind: WorkflowTemplate
metadata:
```

6. Check that you're calling SAP AI Core using the expected tenant.
7. Check if the workflow templates contain the correct scenario label.
8. Get your application sync status using the endpoint:  
**GET** `{{apiurl}}/v2/admin/applications/{{appName}}/status`  
The status will any return errors in your templates. When your templates are updated, the application will resync automatically after approximately three minutes.

## Executables do not appear when you retrieve them from a scenario

### Follow the solution:

1. Delete the synced application using the endpoint:  
**DELETE** `{{apiurl}}/v2/admin/repositories/{{appName}}`
2. Offboard the associated GitHub repository using the endpoint:  
**DELETE** `{{apiurl}}/v2/admin/repositories/{{repositoryName}}`
3. Onboard the associated GitHub repository using a personal access token instead of your GitHub password. For more information, see [Creating a personal access token](#) .

**POST** `{{apiurl}}/v2/admin/repositories`

#### Sample Code

Body:

```
apiVersion: argoproj.io/v1alpha1{
  "name": "aicore-test",
  "url": "https://github.com/john/aicore-test",
  "username": "john",
  "password": "yourGitHubPersonalAccessToken"
}
```

4. Create the application using the endpoint:  
**POST** `{{apiurl}}/v2/admin/applications`
5. Check the ArgoCD applications to determine if the repository has been synchronized correctly for the tenant. For example, check that there are no duplicated workflow names. The value of the name parameter is considered as an executable ID.

```
apiVersion: argoproj.io/v1alpha1
kind: WorkflowTemplate
metadata:
  name: text-clf-train-tutorial
  annotations:
    ...
```

6. Check if the user is calling SAP AI Core using the expected tenant.
7. Check that the scenario label is correct in the workflow templates.
8. Get your application sync status using the endpoint:  
**GET** `{{apiurl}}/v2/admin/applications/{{appName}}/status`  
The status will any return errors in your templates. When your templates are updated, the application will resync automatically after approximately three minutes.

## Application status returns `healthy`, but most other properties are `unknown`

### Follow the solution:

1. Delete the synced application using the endpoint:  
**DELETE** `{{apiurl}}/v2/admin/applications/{{appName}}`
2. Offboard the associated GitHub repository using the endpoint:  
**DELETE**  
`{{apiurl}}/v2/admin/repositories/{{repositoryName}}`

3. Onboard the associated GitHub repository using a personal access token instead of your GitHub password. For more information, see [Creating a personal access token](#).

**POST** `{{apiurl}}/v2/admin/repositories`

#### Sample Code

Body:

```
{
  "name": "aicore-test",
  "url": "https://github.com/john/aicore-test",
  "username": "john",
  "password": "yourGitHubPersonalAccessTokenHere"
}
```

4. Create the application using the endpoint:  
**POST** `{{apiurl}}/v2/admin/applications`
5. Check the ArgoCD applications to determine if the repository has been synchronized correctly for the tenant. For example, check that there are no duplicated workflow names. The value of the name parameter is considered as an executable ID.

```
apiVersion: argoproj.io/v1alpha1
kind: WorkflowTemplate
metadata:
  name: text-clf-train-tutorial
  annotations:
    ...
```

6. Check if the user is calling SAP AI Core using the expected tenant.
7. Check if the workflow templates contain the correct scenario label.
8. Get your application sync status using the endpoint:  
**GET** `{{apiurl}}/v2/admin/applications/{{appName}}/status`  
The status will any return errors in your templates. When your templates are updated, the application will resync automatically after approximately three minutes.

**Application status message:** `rpc error: code = Unknown desc = my-path: app path does not exist`

The specified path in your application doesn't exist in your repository.

#### Follow the solution:

Delete your application and create a new one using the correct path.

**Application status message:** `application repo <your git repository> is not permitted in project 'xyz'`

The repository URL cannot be found in your onboarded repositories.

### Check the following:

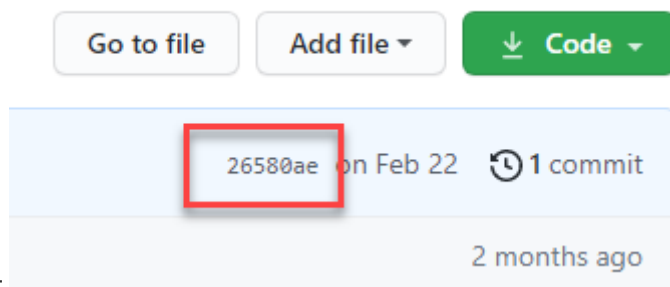
Make sure that the repository specified in your application has been successfully onboarded by using `GET {{apiurl}}/v2/admin/applications` and checking that the repository url has `"status": "COMPLETED"`

**Application status message:** `rpc error: code = Unknown desc = Unable to resolve 'not-existing-branch' to a commit SHA`

The revision you specified in your application doesn't exist in your repository.

### Follow the solution:

Delete your application and create a new one using the correct revision. The revision number is found in GitHub



here:

Alternatively, enter `HEAD` to refer to the latest commit.

**Application status message:** `rpc error: code = FailedPrecondition desc = Failed to unmarshal \"workflow.yaml\": failed to unmarshal manifest: error converting YAML to JSON: yaml: line 7: mapping values are not allowed in this context`

You have a syntax error in your workflow template.

### Follow the solution:

Use Argo Lint to identify syntax errors in your workflow template. To set up the Argo Lint IDE, refer to [Argo Lint IDE setup](#)

**Application status message:** `spec.source.repoURL and spec.source.path either spec.source.chart are required`

You specified an empty path in your application.

```
{  
  "healthStatus": "Unknown",
```

```
"message": "spec.source.repoURL and spec.source.path either
spec.source.chart are required",
"reconciledAt": "Unknown",
"source": {
  "path": "Unknown",
  "repoURL": "Unknown",
  "revision": "Unknown"
},
"syncFinishedAt": "Unknown",
"syncRessourcesStatus": [],
"syncStartedAt": "Unknown",
"syncStatus": "Unknown"
}
```

### Follow the solution:

Delete your application and create a new one, specifying a path. Check its status by using the endpoint:  
`{{apiurl}}/v2/admin/applications/{{appName}}/status`

## Sync an Application Manually

Applications sync with your GitHub repository automatically at intervals of ~3 minutes. Use the endpoint below to manually request a sync:  
`{{apiurl}}/admin/applications/{{appName}}/refresh`

Parent topic: [Troubleshooting \[page 168\]](#)

## Related Information

[Repository \[page 169\]](#)

[Configuration \[page 170\]](#)

[Artifacts \[page 172\]](#)

[Execution \[page 179\]](#)

[Docker \[page 181\]](#)

[Deployment \[page 182\]](#)

[Miscellaneous \[page 183\]](#)

## 13.1.5 Execution

### Execution status is DEAD or PENDING for a long time

#### Check the following:

1. Check the execution logs.
2. Check that the parameter name and execution name that you provided match those in your template. Note that the names are case-sensitive.

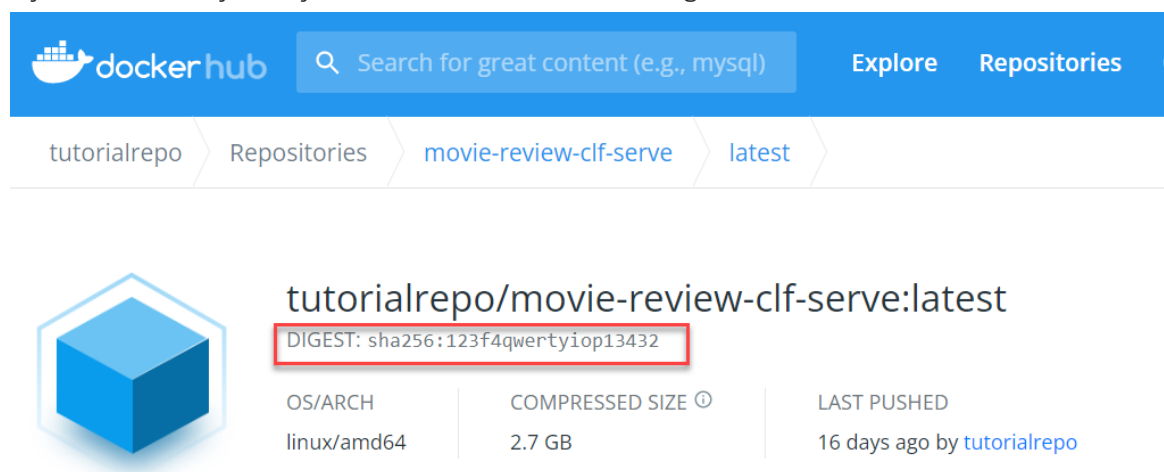
### Images from private docker.io cannot be pulled for execution

#### Follow the solution:

1. Specify the domain name of your Docker Hub. For example:

```
...
        none: {}
    container:
        image: "docker.io/tutorialrepo/movie-review-clf-serve@sha256:123f4qwertyiop13432"
        imagePullPolicy: Always
...
```

If you do not already know your domain name, refer to the image:



The screenshot shows the Docker Hub interface for the image `tutorialrepo/movie-review-clf-serve:latest`. The image ID `DIGEST: sha256:123f4qwertyiop13432` is highlighted with a red box. Below the image ID, the following information is displayed:

| OS/ARCH     | COMPRESSED SIZE ⓘ | LAST PUSHED                 |
|-------------|-------------------|-----------------------------|
| linux/amd64 | 2.7 GB            | 16 days ago by tutorialrepo |

2. Specify your Docker Image digest instead of the image version in your templates:  
image tutorialrepo/movie-review-clf-serve@sha256:123f4qwertyiop13432

## Output artifacts in the execution list are empty and have status UNKNOWN for a long time

### Check the following:

1. Check that an object secret with the name `default` exists in the current resource group. For more information, see [Register an Object Store Secret \[page 102\]](#)
2. Check that you have created the Docker Registry secrets required to pull your Docker Images. Logs will be available only after the execution has started.

## GET execution has the status UNKNOWN for a long time

### Check the following:

1. Check if an object secret with the name `default` exists in the current resource group. For more information, see [Register an Object Store Secret \[page 102\]](#)
2. Check that you have created the Docker Registry secrets required to pull your Docker Images. Logs will be available only after the execution has started.

## Execution status changes to DEAD without any log

### Check the following:

1. Check that your templates meet the Argo specifications and that they can be executed by SAP AI Core.
2. To identify errors automatically in your templates, use the Argo linter.
3. Check that you have created Docker folders, to store artifacts that will be created during an execution.

Parent topic: [Troubleshooting \[page 168\]](#)

## Related Information

[Repository \[page 169\]](#)

[Configuration \[page 170\]](#)

[Artifacts \[page 172\]](#)

[Application \[page 174\]](#)

[Docker \[page 181\]](#)

[Deployment \[page 182\]](#)

[Miscellaneous \[page 183\]](#)

## 13.1.6 Docker

### Error pulling container main logs. Back-off pulling image

You push a template to SAP AI Core and you get the error "Error pulling container main logs".  
"Back-off pulling image".

#### Complete the following steps:

1. Make sure that your Docker Registry is public-facing and not protected behind the firewall of your organization.  
If your Docker Image isn't public, verify that you have created a Docker Registry secret in SAP AI Core.
2. Check that you can pull your Docker Image on your local computer using the credentials from the previous step.
3. Specify the same Docker Registry secret in your executable.
4. Specify the path to your Docker Image in your executable in following format:  
<DOCKER\_REGISTRY> / <REPO\_NAME> / <DOCKER\_IMAGE> : <TAGNAME>

#### ❁ Example

```
docker.io/tutorialrepo/text-clf-train:0.0.1
```

Parent topic: [Troubleshooting \[page 168\]](#)

### Related Information

[Repository \[page 169\]](#)

[Configuration \[page 170\]](#)

[Artifacts \[page 172\]](#)

[Application \[page 174\]](#)

[Execution \[page 179\]](#)

[Deployment \[page 182\]](#)

[Miscellaneous \[page 183\]](#)

## 13.1.7 Deployment

### You want to force a deployment with status UNKNOWN to stop

You want to stop and delete a deployment but you are unable to because the deployment status is “Unknown”. You have tried to submit a PATCH request as follows:

```
PATCH {{apiurl}}/lm/deployments/d4fec9c24c54f87e
```

However, you receive the following response:

```
{
  "error": {
    "code": "01010076"
    "message": "Invalid Request, Current status UNKNOWN cannot be changed..",
    "requestID": "e110820e-1cfe-456a-bb0e-77907b36422c",
    "target": "/apu/v2/deployments/d4fec9c24c54f87e"
  }
}
```

#### Complete the following steps:

1. Find out why your deployment status is “Unknown” by using the endpoint:  
`GET $AI_API_URL/v2/lm/deployments/<deploymentid>`
2. Delete the deployment without trying to stop it (stopping a deployment is necessary only when it's running):  
`DELETE $AI_API_URL/v2/lm/deployments/<deploymentid>`

### Deployment remains in status PENDING

#### Check the following:

1. Check that the Docker Registry secret exists when using your private Docker Image.
2. Check that your Docker Image can be downloaded to your local system.

### Deployment ID <abc> not found

This message appears when you have just started the deployment. Wait a few minutes and the message will resolve itself automatically.

Parent topic: [Troubleshooting \[page 168\]](#)

## Related Information

[Repository \[page 169\]](#)

[Configuration \[page 170\]](#)

[Artifacts \[page 172\]](#)

[Application \[page 174\]](#)

[Execution \[page 179\]](#)

[Docker \[page 181\]](#)

[Miscellaneous \[page 183\]](#)

## 13.1.8 Miscellaneous

### 403 - forbidden: RBAC Access denied

When you submit a POST request for an execution or a configuration, you get the error 403 - forbidden: RBAC Access denied.

#### Check the following:

1. Check that you are passing the correct token and AI-Resource-Group header.
2. Check your tenant provisioning.

### You get a runtime adapter exception

```
You get the error: "Runtime Adapter Exception; Failed to post deployments : {\n\n  \"code\": \"400\", \n  \"message\": \"Missing input parameter or artifacts, one or\n  more placeholder values are not resolved in the serving spec and error is 'dict\n  object' has no attribute '.'.\\n}\\n"
```

#### Check the following:

Check that your input artifact key doesn't contain any separators, such as "example-artifact-key". If it does, rename the key (for example, "exampleArtifactKey")

## You have pushed your template to your GitHub repository but you can't see the executable created via the API

### There may be an error in your template. Check the following:

- Serving templates always expect an input parameter to be configured. If you don't have a parameter in your Serving template, add a dummy parameter and create a configuration for it.
- Hyphens (-) are the only separator as permitted in the template name.

## You have created a scenario in a template but you cannot see it in the AI API calls

### Use the following solution:

Add a workflow template with the new `scenario_id` (not a serving template) to make the scenario visible.

## Error: getaddrinfo ENOTFOUND

### Check the following:

1. Check that all environment variables match your SAP AI Core service keys. Specifically, check:
  - `auth_url`
  - `client_id`
  - `client_secret`
  - `apiurl`
2. Submit a GET request to the endpoint `{{apiurl}}/v2/admin/repositories`
3. If the issue persists, contact SAP support as described at [Monitoring and Troubleshooting \[page 168\]](#).

## Git repo doesn't synchronize with SAP AI Core instances.

When you try to synchronize your git repository with SAP AI Core, the response shows empty fields.

### Check the following:

1. Check that you are using a GitHub personal access token and not your GitHub password.

If you are already using a personal access token, proceed as follows:

1. Delete all UNKNOWN applications from your SAP AI Core instance using the endpoint:  
`DELETE {{apiurl}}/v2/admin/applications/{{appName}}`
2. Offboard your GitHub repo from SAP AI Core by calling the endpoint:  
`DELETE {{apiurl}}/v2/admin/repositories/{{repositoryName}}`
3. Generate a GitHub personal access token and use it to onboard your git repo to SAP AI Core as before. For more information, see [Creating a personal access token](#) .

4. Sync your application again.

**Parent topic:** [Troubleshooting \[page 168\]](#)

## **Related Information**

[Repository \[page 169\]](#)

[Configuration \[page 170\]](#)

[Artifacts \[page 172\]](#)

[Application \[page 174\]](#)

[Execution \[page 179\]](#)

[Docker \[page 181\]](#)

[Deployment \[page 182\]](#)

# 14 Support Process

Explore solutions to potential issues, and find out how to get support.

## Getting Support

If you encounter an issue with this service, we recommend that you follow the procedure below.

### Check Platform Status

Check the availability of the platform at [SAP Trust Center](#).

For more information about platform availability, updates and notifications, see [Platform Updates and Notifications in the Cloud Foundry Environment](#).

### Check Guided Answers

In the SAP Support Portal, check the [Guided Answers](#) section for SAP Business Technology Platform. You can find solutions for general SAP Cloud Platform issues as well as for specific services there.

### Contact SAP Support

You can report an incident or error through the [SAP Support Portal](#).

Please use the following component(s) for your incident:

| Component Name | Component Description |
|----------------|-----------------------|
| CA-ML-AIC      | SAP AI Core           |

We recommend that you include the following information when you submit the incident:

- Region information
- Subaccount technical name
- URL of the page where the incident or error occurs
- Steps or clicks used to replicate the error
- Screenshots, videos, or the code entered

# 15 Service Offboarding

Tenant offboarding occurs when a customer deletes a subaccount. SAP AI Core polls for the subaccount deletion event and performs the necessary deprovisioning and deletion activities.

## ⓘ Note

Data and resources are not deleted when a service instance is deleted (because we don't isolate based on the service instance). If you want to keep the subaccount but still deprovision SAP AI Core, create a **medium support ticket** on component `CA-ML-AIC` with the title **Service Offboarding** and request that your data and resources be deleted manually.

## → Remember



When you delete a resource group, [Manage mTLS Certificate Secrets \[page 130\]](#) in that group are deleted. Certificates that were issued before the deletion may remain valid until they expire. You are responsible for removing or revoking trust on any external services that rely on those certificates.

# 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 an 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.

## Bias-Free Language

SAP supports a culture of diversity and inclusion. Whenever possible, we use unbiased language in our documentation to refer to people of all cultures, ethnicities, genders, and abilities.



© 2026 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.

