Workload YAML Reference

Version Control

Version #	What's new?	Date updated
1	New version	June 2024

Table of contents

Workload YAML Reference

Each field has a specific value type. The following value types are supported.

Value types

Value type	Description	Example
`Boolean`	A binary value that can be either True or False	`true`
`String`	A sequence of characters used to represent text. It can include letters, numbers, symbols, and spaces	`ubuntu:latest`

`Itemized`	An ordered collection of items (objects), which can be of different types (all items in the list are of the same type).	`pvcs:`
		` items:`
		` pvc-1:`
		`  value:`
		`  claimName: pvc-1`
		`  ephemeral: false`
		`  path: /tmp/pvc-1`
		`  readOnly: false`
		`  readWriteOnce: true`
		`  size: 1G`
		`  storageClass: default`
		` pvc-2:`
		`  value:`
		`  claimName: pvc-2`
		`  ephemeral: false`
		`  path: /tmp/pvc-2`
		`  readOnly: false`
		`  readWriteOnce: true`
		`  size: 1G`
		`  storageClass: default`

`Integer`	An Integer is a whole number without a fractional component.	`4`
`Number`	Capable of having non-integer values	`5.5 or 0.3 etc`
`Quantity`	Holds a string composed of a number and a unit representing a quantity	`10G or 15M etc`

Workload YAML reference table

The fields listed in the table are all the fields below the spec (spec.field).
Click the link to view full details of each field.

Fields	Description	Supported Run:ai workload types
allowPrivilegeEscalation	Allows the container running the workload and all launched processes to gain additional privileges after the workload starts	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
args	When set, contains the arguments sent along with the command. These override the entry point of the image in the created workload.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
capabilities	The capabilities field allows adding a set of unix capabilities to the container running the workload. Capabilities are Linux distinct privileges traditionally associated with superuser which can be independently enabled and disabled	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
command	A command to serve as the entry point of the container running the workspace	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
createHomeDir	Instructs the system to create a temporary home directory for the user within the container. Data stored in this directory is not saved when the container exits. When the runAsUser flag is set to true, this flag defaults to true as well.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
environment	Set of environmentVariables to populate the container running the workspace	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)

Fields	Description	Supported Run:ai workload types
hostipc	Whether to enable hostIpc. Defaults to false.	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
hostNetwork	Whether to enable host network.	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
image	Specifies the image to use when creating the container running the workload.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
imagePullPolicy	Specifies the pull policy of the image when starting a container running the created workload. Options are: always, ifNotPresent, or never.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
runAsUid	Specifies the Unix user id with which the container running the created workload should run. It is used only if runAsUser is set to true.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
runAsGid	Specifies the Unix Group ID with which the container should run. It is used only if runAsUser is set to true.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
supplementalGroups	Comma separated list of groups that the user running the container belongs to, in addition to the group indicated by runAsGid	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)

Fields	Description	Supported Run:ai workload types
workingDir	Container's working directory. If not specified, the container runtime default is used, which might be configured in the container image.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
runAsNonRoot	Indicates that the container must run as a non-root user	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
seccompProfileType	Indicates which kind of seccomp profile is applied to the container. The options are: ● RuntimeDefault - the container runtime default profile should be used ● Unconfined - no profile should be applied	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
extendedResources	Specifies values for extended resources. Extended resources are third-party devices (such as high-performance NICs, FPGAs, or InfiniBand adapters) that you want to allocate to your Job.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
migProfile	Specifies the memory profile to be used for workload running on NVIDIA Multi-Instance GPU (MIG) technology.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
nodeType	Nodes (machines) or a group of nodes on which the workload runs.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
nodePools	A prioritized list of node pools for the scheduler to run the workspace on. The scheduler always tries to use the first node pool before moving to the next one when the first is not available.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)

Fields	Description	Supported Run:ai workload types
annotations	Set of annotations to populate into the container running the workspace	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
labels	Set of labels to populate into the container running the workspace	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
terminateAfterPreemption	Indicates whether the job should be terminated, by the system, after it has been preempted	☑ Workspace ☑ Training ☐ Distributed ☐ Inference ☑ Job (legacy)
autoDeletionTimeAfterCompletionSeconds	Specifies the duration after which a finished workload (Completed or Failed) is automatically deleted. If this field is set to zero, the workload becomes eligible to be deleted immediately after it finishes.	☑ Workspace ☑ Training ☐ Distributed ☐ Inference ☐ Job (legacy)
backoffLimit	Specifies the number of retries before marking a workload as failed	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
completions	Used with Hyperparameter Optimization. Specifies the number of successful pods the job should reach to be completed. The Job is marked as successful once the specified amount of pods has succeeded	☐ Workspace ☑ Training ☐ Distributed ☐ Inference ☑ Job (legacy)
Parallelism	Used with Hyperparameter Optimization. Specifies the maximum desired number of pods the workload should run at any given time.	☐ Workspace ☑ Training ☐ Distributed ☐ Inference ☑ Job (legacy)

Fields	Description	Supported Run:ai workload types
gitSync	Details of the git repository and items mapped to it	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
pvcs	Specifies persistent volume claims to mount into a container running the created workload	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
nfs	Specifies NFS volume to mount into the container running the workload	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
s3	Specifies S3 buckets to mount into the container running the workload	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
configMapVolumes	Specifies ConfigMaps to mount as volumes into a container running the created workload.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
active	Specifies whether the workload should be active or suspended	☑ Workspace ☑ Training ☐ Distributed ☐ Inference ☑ Job (legacy)
cpu	Specifies CPU units to allocate for the created workload (0.5, 1, .etc). The workload receives at least this amount of CPU. Note that the workload is not scheduled unless the system can guarantee this amount of CPUs to the workload.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)

Fields	Description	Supported Run:ai workload types
cpuLimit	Specifies a limit on the number of CPUs consumed by the workload (0.5, 1, .etc). The system guarantees that this workload is not able to consume more than this amount of CPUs.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
dropAllCapabilities	Indicates whether to drop all capabilities or not. Default is false	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
exposedUrls	Specifies a set of exported URL (e.g. ingress) from the container running the created workload.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)
gpu	Specifies the fraction of GPU to be allocated to the workload, between 0 and 1. For backward compatibility, it also supports the number of gpuDevices larger than 1, currently provided using the gpuDevices field.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
gpuDevices	Specifies the number of GPUs to allocate for the created workload	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
gpuLimit	When a fraction of a GPU is requested, the GPU limit specifies the portion limit to allocate to the workload. The range of the value is from 0 to 1.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
gpuMemory	Specifies GPU memory to allocate for the created workload. The workload receives this amount of memory. Note that the workload is not scheduled unless the system can guarantee this amount of GPU memory to the workload.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)

Fields	Description	Supported Run:ai workload types
gpuMemoryLimit	Specifies a limit on the GPU memory to allocate for this workload. Should be no less than the gpuMemory.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
largeShm	Specifies a large /dev/shm device to mount into a container running the created workload. SHM is a shared file system mounted on RAM.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
memory	Specifies the amount of CPU memory to allocate for this workload (1G, 20M, .etc). The workload receives at least this amount of memory. Note that the workload is not scheduled unless the system can guarantee this amount of memory to the workload.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
memoryLimit	Specifies a limit on the CPU memory to allocate for this workload (1G, 20M, .etc). The system guarantees that this workload is not able to consume more than this amount of memory. The workload receives an error when trying to allocate more memory than this limit.largeShmRequest	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
mountPropagation	Allows for sharing volumes mounted by a container to other containers in the same pod, or even to other pods on the same node. The volume mount receives all subsequent mounts that are mounted to this volume or any of its subdirectories.	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
name	The specific name of the created workload. Either name of namePrefix should be provided, but not both.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
namePrefix	A prefix used for assigning a name to the created resource. Either name of namePrefix should be provided, but not both.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)

Fields	Description	Supported Run:ai workload types
podAffinity	Indicates whether pod affinity scheduling rules apply	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
podAffinitySchedulingRule	Indicates if we want to use the Pod affinity rule as : the "hard" (required) or the "soft" (preferred) This field can be specified only if PodAffinity is set to true	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
podAffinityTopology	Specifies the Pod Affinity Topology to be used for scheduling the job. This field can be specified only if PodAffinity is set to true	☑ Workspace ☑ Training ☐ Distributed ☑ Inference ☑ Job (legacy)
ports	Specifies a set of ports exposed from the container running the created workload. Used together with --service-type.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
preemptible	Specifies that the created workload is preemptible. Interactive preemptible workloads can be scheduled above the guaranteed quota but may be reclaimed at any time.	☑ Workspace ☐ Training ☐ Distributed ☐ Inference ☐ Job (legacy)
processes	Number of distributed training processes that are allocated for the created mpijob (MPI ONLY)	☐ Workspace ☐ Training ☑ Distributed ☐ Inference ☐ Job (legacy)
readOnlyRootFilesystem	If true, mounts the container's root filesystem as read-only	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)

Fields	Description	Supported Run:ai workload types
readinessProbe	Specifies the ReadinessProbe to use to determine if the container is ready to accept traffic. For more information see https://kubernetes.io/docs/tasks/configure-pod-container/configure-liveness-readiness-startup-probes	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
runAsUser	Limits the container running the created workload to run in the context of a specific non-root user. The user id is provided by the runAsUid field. This would manifest itself in access to operating system resources, in the ownership of new folders created under shared directories, etc. Alternatively, if your cluster is connected to Run:ai via SAML, you can map the container to use the Linux UID/GID which is stored in the organization's directory. For more information see the User Identity guide at https://docs.run.ai/admin/runai-setup/config/non-root-containers/	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)
serviceType	Specifies the default service exposure method for ports. The default shall be used for ports which do not specify service type. Options are: LoadBalancer, NodePort or ClusterIP. For more information see the External Access to Containers guide on https://docs.run.ai/admin/runai-setup/config/allow-external-access-to-containers/	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)
slotsPerWorker	Number of slots to allocate per worker in the created mpijob (MPI ONLY)	☐ Workspace ☐ Training ☑ Distributed ☐ Inference ☐ Job (legacy)
stdin	Instructs the system to keep stdin open for the container(s) running the created workload, even if nothing is attached	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)

Fields	Description	Supported Run:ai workload types
tolerations	Toleration rules which apply to the pods running the workload. Toleration rules guide (but do not require) the system to which node each pod can be scheduled to or evicted from, based on matching between those rules and the set of taints defined for each Kubernetes node.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
tty	Instructs the system to allocate a pseudo-TTY for the created workload.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)
username	Display-only field describing the user who owns the workload. The data is not used for authentication or authorization purposes.	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)
volumes	Specifies volumes to mount into a container running the created workload	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)
minScale	The minimum number of replicas to run	☐ Workspace ☐ Training ☐ Distributed ☑ Inference ☐ Job (legacy)
maxScale	The maximum number of replicas to run	☐ Workspace ☐ Training ☐ Distributed ☑ Inference ☐ Job (legacy)
metric	The predefined metric to use for autoscaling. Possible values are: throughput, concurrency and latency	☐ Workspace ☐ Training ☐ Distributed ☑ Inference ☐ Job (legacy)

Fields	Description	Supported Run:ai workload types
target	The target value for the autoscaling metric	☐ Workspace ☐ Training ☐ Distributed ☑ Inference ☐ Job (legacy)

YAML field details

Field	`allowPrivilegeEscalation`		
Description	Allows the container running the workload and all launched processes to gain additional privileges after the workload starts		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	Type `Boolean`	Mandatory `No`	Default value `True`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: ubuntu:latest
	  allowPrivilegeEscalation:
	  value: true
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAML reference table

Field	`args`		
Description	When set, contains the arguments sent along with the command. These override the entry point of the image in the created workload.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: jupyter/scipy-notebook
	  command:
	  value: start-notebook.sh
	  arguments:
	  value:
	--NotebookApp.base_url=/${RUNAI_PROJECT}/${RUNAI_JOB_NAME} --NotebookApp.token=''
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAMl reference table

Field	`capabilities`		
Description	The capabilities field allows adding a set of unix capabilities to the container running the workload. Capabilities are Linux distinct privileges traditionally associated with superuser which can be independently enabled and disabled		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

```
apiVersion: run.ai/v2alpha1
kind: InteractiveWorkload
metadata:
  name: <name>
  namespace: <namespace>
spec:
  image:
    value: ubuntu:latest
  capabilities:
    items:
      "cap-1":
        value:
          capability: AUDIT_CONTROL
      "cap-2":
        value:
          capability: IPC_LOCK
  name:
    value: <workload-name>
  nodePools:
    value: default
  usage: Submit
```

Return to Workload YAMl reference table

Field	`command`		
Description	A command to serve as the entry point of the container running the workspace		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: jupyter/scipy-notebook
	  command:
	  value: start-notebook.sh
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAMl reference table

Field	`createHomeDir`		
Description	Instructs the system to create a temporary home directory for the user within the container. Data stored in this directory is not saved when the container exits. When the runAsUser flag is set to true, this flag defaults to true as well.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `True`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: ubuntu:latest
	  createHomeDir:
	  value: true
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAMl reference table

Field	`environment`		
Description	Set of environment variables to populate the container running the workspace		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Map (key → value)`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`hostipc`		
Description	Whether to enable hostIpc. Defaults to false.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `False`
Example workload snippet			

Return to Workload YAML reference table

Field	`hostNetwork`		
Description	Whether to enable host networking.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `False`
Example workload snippet			

Return to Workload YAML reference table

Field	`image`		
Description	Specifies the image to use when creating the container running the workload.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `Yes`	**Default value** `N/A`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: ubuntu:latest
	  imagePullPolicy:
	  value: Always
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAMl reference table

Field	`imagePullPolicy`		
Description	Specifies the pull policy of the image when starting a container running the created workload. Options are: always, ifNotPresent, or never.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `Always`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: ubuntu:latest
	  imagePullPolicy:
	  value: Always
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAML reference table

Field	`runAsUid`		
Description	Specifies the Unix user id with which the container running the created workload should run. Is used only if runAsUser is set to true.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Integer`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`runAsGid`		
Description	Specifies the Unix Group ID with which the container should run. Is used only if runAsUser is set to true.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Integer`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAMl reference table

Field	`supplementalGroups`		
Description	Comma separated list of groups that the user running the container belongs to, in addition to the group indicated by runAsGid		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAMl reference table

Field	`workingDir`		
Description	Container's working directory. If not specified, the container runtime default is used, which might be configured in the container image.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: ubuntu:latest
	  workingDir:
	  value: /tmp
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAML reference table

Field	`runAsNonRoot`		
Description	Indicates that the container must run as a non-root user		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `Yes`	**Default value** `True`
Example workload snippet			

Return to Workload YAML reference table

Field	`seccompProfileType`		
Description	Indicates which kind of seccomp profile is applied to the container. The options are ● RuntimeDefault - the container runtime default profile should be used ● Unconfined - no profile should be applied		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`extendedResources`		
Description	Specifies values for extended resources. Extended resources are third-party devices (such as high-performance NICs, FPGAs, or InfiniBand adapters) that you want to allocate to your Job.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Map (Key → Value)`	**Mandatory** `No`	**Default value** `N/A`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: ubuntu:latest
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit
	  extendedResources:
	  items:
	  "resrc-1":
	  value:
	  name: "ext/a"
	  quantity: "5"
	  "resrc-2":
	  value:
	  name: "ext/b"
	  quantity: "10M"</pre>		

Return to Workload YAMl reference table

Field	`migProfile`		
Description	Specifies the memory profile to be used for workload running on NVIDIA Multi-Instance GPU (MIG) technology.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`nodeType`		
Description	Nodes (machines) or a group of nodes on which the workload runs.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAMl reference table

Field	`nodePools`		
Description	A prioritized list of node pools for the scheduler to run the workspace on. The scheduler always tries to use the first node pool before moving to the next one when the first is not available.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`annotations`		
Description	Set of annotations to populate into the container running the workspace		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Map (key → value)`	**Mandatory** `No`	**Default value** `N/A`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  annotations:
	  items:
	  id:
	  value: b48a92e1-02e3-4592-acb2-59a4b027cc0b
	  image:
	  value: ubuntu:latest
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAML reference table

Field	`labels`		
Description	Set of labels to populate into the container running the workspace		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Map (key → value)`	**Mandatory** `No`	**Default value** `N/A`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  labels:
	  items:
	  PROJECT:
	  value: TEST
	  image:
	  value: ubuntu:latest
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAML reference table

Field	`terminateAfterPreemption`		
Description	Indicates whether the job should be terminated by the system, after it has been preempted		
Supported Run:ai workload types	☑ Workspace ☑ Training ☐ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `False`
Example workload snippet			

Return to Workload YAML reference table

Field	`autoDeletionTimeAfterCompletionSeconds`		
Description	Specifies the duration after which a finished workload (Completed or Failed) is automatically deleted. If this field is set to zero, the workload becomes eligible to be deleted immediately after it finishes.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☐ Distributed ☐ Inference ☐ Job (legacy)		
Value	**Type** `Integer`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`backoffLimit`		
Description	Specifies the number of retries before marking a workload as failed		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Integer`	**Mandatory** `No`	**Default value** `6`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: ubuntu:latest
	  backoffLimit:
	  value: 6
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAMl reference table

Field	`completions`		
Description	Used with Hyperparameter Optimization. Specifies the number of successful pods the job should reach to be completed. The Job is marked as successful once the specified amount of pods has succeeded		
Supported Run:ai workload types	☐ Workspace ☑ Training ☐ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Integer`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`parallelism`		
Description	Used with Hyperparameter Optimization. Specifies the maximum desired number of pods the workload should run at any given time.		
Supported Run:ai workload types	☐ Workspace ☑ Training ☐ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Integer`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Spec Fields	`gitSync`		
Description	Details of the git repository and items mapped to it		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	**Default value** `N/A`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: ubuntu:latest
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit
	  gitSync:
	  items:
	  git-kubernetes:
	  value:
	  directory: /tmp/
	  repository: https://github.com/kubernetes/kubernetes
	  rev: release-1.7</pre>		

Spec git fields	**Description**	**Value type**
`Repository` (mandatory)	URL to a remote git repository. The content of this repository is mapped to the container running the workload	`String`
`rev`	Specific revision to synchronize the repository from	`String`
`Path` (mandatory)	Local path within the workspace to which the S3 bucket is mapped	`String`

`secretName`	Optional name of Kubernetes secret that holds your git username and password	`String`
`username`	If secretName is provided, this field should contain the key, within the provided Kubernetes secret, which holds the value of your git username. Otherwise, this field should specify your git username in plain text (example: myuser).	`String`

Return to Workload YAML reference table

Spec fields	`pvcs`		
Description	Specifies persistent volume claims to mount into a container running the created workload		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	**Default value** `N/A`

Example workload snippet	`apiVersion: run.ai/v2alpha1`
	`kind: InteractiveWorkload`
	`metadata:`
	`  name: <name>`
	`  namespace: <namespace>`
	`spec:`
	`  image:`
	`  value: ubuntu:latest`
	`  name:`
	`  value: <workload-name>`
	`  nodePools:`
	`  value: default`
	`  usage: Submit`
	`  pvcs:`
	`  items:`
	`  pvc-1:`
	`  value:`
	`  claimName: pvc-1`
	`  ephemeral: false`
	`  path: /tmp/`
	`  readOnly: false`
	`  readWriteOnce: true`
	`  size: 1G`
	`  storageClass: default`		

Spec PVC fields	**Description**	**Value type**
`claimName` (mandatory)	A given name for the PVC. Allowed referencing it across workspaces	`String`
`ephemeral`		`Boolean`
`path`		

`readOnly`	Permits read only from the PVC, prevents additions or modifications to its content	`Boolean`
`ReadwriteOnce`	Requesting claim that can be mounted in read/write mode to exactly 1 host. If none of the modes are specified, the default is readWriteOnce	`Boolean`
`Size`	Requested size for the PVC. Mandatory when existing PVC is false	`String`
`storageClass`	Storage class name to associate with the PVC. This parameter may be omitted if there is a single storage class in the system, or you are using the default storage class. Further details at https://kubernetes.io/docs/concepts/storage/storage-classes.	`String`
`readOnlyMany`	Requesting claim that can be mounted in read-only mode to many hosts	`Boolean`
`readWriteMany`	Requesting claim that can be mounted in read/write mode to many hosts	`Boolean`

Return to Workload YAML reference table

Spec fields	`nfs`		
Description	Specifies NFS volume to mount into the container running the workload		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	Default Value `N/A`
Example workload snippet			

Field	Description	Value type
`sourcePath` (mandatory)	The path that the NFS volume is mounted to when in use	`String`
`readOnly`	Whether to force the NFS export to be mounted with read-only permissions	`Boolean`
`nfsServer` (mandatory)	The hostname or IP address of the NFS server	`String`
`targetPath` (mandatory)	The path that the NFS volume is mounted to when in use	`String`

Return to Workload YAML reference table

Field	`s3`		
Description	Specifies S3 buckets to mount into the container running the workload)		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Field	**Description**	**Value type**
`Bucket` (mandatory)	The name of the bucket	`String`
`Path` (mandatory)	Local path within the workspace to which the S3 bucket is mapped	`String`
`url`	The URL of the S3 service provider. The default is the URL of the Amazon AWS Se service	`String`
`key`	If secretName is provided, this field should contain the key, within the provided kubernetes secret, which holds the value of your S3 access key ID. otherwise, this field should specify your S3 access key ID in plain text (example: AKIAIOSFODNN7EXAMPLE).	`String`
`secret`	If secretName is provided, this field should contain the key, within the provided kubernetes secret, which holds the value of your S3 secret access key. otherwise, this field should specify your S3 secret access key in plain text (example: wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY).	`String`
`secretName`	Optional name of kubernetes secret that holds your S3 access key ID and S3	`String`

Return to Workload YAML reference table

Field	`configMapVolume`		
Description	Specifies ConfigMaps to mount as volumes into a container running the created workload.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	**Default value** `N/A`

Example workload snippet	<pre>apiVersion: run.ai/v2alpha1
	kind: InteractiveWorkload
	metadata:
	  name: <name>
	  namespace: <namespace>
	spec:
	  image:
	  value: ubuntu:latest
	  configMapVolumes:
	  items:
	  configmap-volume-1:
	  value:
	  name: cm-db-setting
	  mountPath: /etc/db.conf
	  name:
	  value: <workload-name>
	  nodePools:
	  value: default
	  usage: Submit</pre>		

Return to Workload YAML reference table

Field	`active`		
Description	Specifies whether the workload should be active or suspended		
Supported Run:ai workload types	☑ Workspace ☑ Training ☐ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `N/A`	**Default value** `True`
Example workload snippet			

Return to Workload YAMl reference table

Field	`cpu`		
Description	Specifies CPU units to allocate for the created workload (0.5, 1, .etc). The workload receives at least this amount of CPU. Note that the workload is not scheduled unless the system can guarantee this amount of CPUs to the workload.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Number`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`cpuLimit`		
Description	Specifies a limit on the number of CPUs consumed by the workload (0.5, 1, etc). The system guarantees that this workload is not able to consume more than this amount of CPUs.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Number`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`dropAllCapabilities`		
Description	Indicates whether to drop all capabilities or not. Default is false.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `False`
Example workload snippet			

Return to Workload YAMl reference table

Field	`exposedUrls`		
Description	Specifies a set of exported URL (e.g. ingress) from the container running the created workload		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`gpu`		
Description	Specifies the fraction of GPU to be allocated to the workload, between 0 and 1. For backward compatibility, it also supports the number of gpuDevices larger than 1, currently provided using the gpuDevices field.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Number`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`gpuDevices`		
Description	Specifies the number of GPUs to allocate for the created workload		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Integer`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAMl reference table

Field	`gpuLimit`		
Description	When a fraction of a GPU is requested, the GPU limit specifies the portion limit to allocate to the workload. The range of the value is from 0 to 1.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Number`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAMl reference table

Field	`gpuMemory`		
Description	Specifies GPU memory to allocate for the created workload. The workload receives this amount of memory. Note that the workload is not scheduled unless the system can guarantee this amount of GPU memory to the workload.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Quantity`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`gpuMemoryLimit`		
Description	Specifies a limit on the GPU memory to allocate for this workload. Should be no less than the gpuMemory.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Quantity`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`largeShm`		
Description	Specifies a large /dev/shm device to mount into a container running the created workload. SHM is a shared file system mounted on RAM.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAMl reference table

Field	`memory`		
Description	Specifies the amount of CPU memory to allocate for this workload (1G, 20M, .etc). The workload receives at least this amount of memory. Note that the workload is not scheduled unless the system can guarantee this amount of memory to the workload		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Quantity`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`memoryLimit`		
Description	Specifies a limit on the CPU memory to allocate for this workload (1G, 20M, etc). The system guarantees that this workload is not able to consume more than this amount of memory. The workload receives an error when trying to allocate more memory than this limit.largeShmRequest		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Quantity`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`mountPropagation`		
Description	Allows for sharing volumes mounted by a container to other containers in the same pod, or even to other pods on the same node. The volume mount receives all subsequent mounts that are mounted to this volume or any of its subdirectories.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`name`		
Description	The specific name of the created workload. Either name of namePrefix should be provided, but not both.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`namePrefix`		
Description	A prefix used for assigning a name to the created resource. Either name of namePrefix should be provided, but not both.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`podAffinity`		
Description	Indicates whether pod affinity scheduling rules apply		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`podAffinitySchedulingRule`		
Description	Indicates if we want to use the Pod affinity rule as : the "hard" (required) or the "soft" (preferred) This field can be specified only if PodAffinity is set to true		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`podAffinityTopology`		
Description	Specifies the Pod Affinity Topology to be used for scheduling the job. This field can be specified only if PodAffinity is set to true.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☐ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`ports`		
Description	Specifies a set of ports exposed from the container running the created workload. Used together with `--service-type`.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`preemptible`		
Description	Specifies that the created workload is preemptible. Interactive preemptible workloads can be scheduled above the guaranteed quota but may be reclaimed at any time.		
Supported Run:ai workload types	☑ Workspace ☐ Training ☐ Distributed ☐ Inference ☐ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`processes`		
Description	Number of distributed training processes that are allocated for the created mpijob		
Supported Run:ai workload types	☐ Workspace ☐ Training ☑ Distributed ☐ Inference ☐ Job (legacy)		
Value	**Type** `Integer`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAMl reference table

Field	`readOnlyRootFilesystem`		
Description	If true, mounts the container's root filesystem as read-only		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `False`
Example workload snippet			

Return to Workload YAMl reference table

Field	`readinessProbe`		
Description	Specifies the ReadinessProbe to use in order to determine if the container is ready to accept traffic. For more information see https://kubernetes.io/docs/tasks/configure-pod-container/configure-liveness-readiness-startup-probes		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`runAsUser`		
Description	Limits the container running the created workload to run in the context of a specific non-root user. The user id is provided by the runAsUid field. This would manifest itself in access to operating system resources, in the ownership of new folders created under shared directories, etc. Alternatively, if your cluster is connected to Run:ai via SAML, you can map the container to use the Linux UID/GID which is stored in the organization's directory. For more information see the User Identity guide at https://docs.run.ai/admin/runai-setup/config/non-root-containers		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`serviceType`		
Description	Specifies the default service exposure method for ports. The default shall be used for ports which do not specify service type. Options are: LoadBalancer, NodePort or ClusterIP. For more information see the External Access to Containers guide on https://docs.run.ai/admin/runai-setup/config/allow-external-access-to-containers/		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`slotsPerWorker`		
Description	Number of slots to allocate per worker in the created mpijob		
Supported Run:ai workload types	☐ Workspace ☐ Training ☑ Distributed ☐ Inference ☐ Job (legacy)		
Value	**Type** `Integer`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAMl reference table

Field	`stdin`		
Description	Instructs the system to keep stdin open for the container(s) running the created workload, even if nothing is attached		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`tolerations`		
Description	Toleration rules which apply to the pods running the workload. Toleration rules guide (but do not require) the system to which node each pod can be scheduled to or evicted from, based on matching between those rules and the set of taints defined for each Kubernetes node.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`tty`		
Description	Instructs the system to allocate a pseudo-TTY for the created workload.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☑ Job (legacy)		
Value	**Type** `Boolean`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`username`		
Description	Display-only field describing the user who owns the workload. The data is not used for authentication or authorization purposes.		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☑ Inference ☐ Job (legacy)		
Value	**Type** `String`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table

Field	`volumes`		
Description	Specifies volumes to mount into a container running the created workload		
Supported Run:ai workload types	☑ Workspace ☑ Training ☑ Distributed ☐ Inference ☑ Job (legacy)		
Value	**Type** `Itemized`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAMl reference table

Field	`minScale`		
Description	The minimum number of replicas to run		
Supported Run:ai workload types	☐ Workspace ☐ Training ☐ Distributed ☑ Inference ☐ Job (legacy)		
Value	**Type** `Object`	**Mandatory** `No`	**Default value** `0 if metric is concurrency or throughput, 1 otherwise`
Example workload snippet			

Return to Workload YAML reference table

Field	`maxScale`		
Description	The maximum number of replicas to run		
Supported Run:ai workload types	☐ Workspace ☐ Training ☐ Distributed ☑ Inference ☐ Job (legacy)		
Value	**Type** `Object`	**Mandatory** `No`	**Default value** `1`
Example workload snippet			

Return to Workload YAMl reference table

Field	`metric`		
Description	The predefined metric to use for autoscaling. Possible values are throughput, concurrency or latency		
Supported Run:ai workload types	☐ Workspace ☐ Training ☐ Distributed ☑ Inference ☐ Job (legacy)		
Value	**Type** `Object`	**Mandatory** `No`	**Default value** `concurrency`
Example workload snippet			

Return to Workload YAML reference table

Field	`target`		
Description	The target value for autoscaling metric		
Supported Run:ai workload types	☐ Workspace ☐ Training ☐ Distributed ☑ Inference ☐ Job (legacy)		
Value	**Type** `Object`	**Mandatory** `No`	**Default value** `N/A`
Example workload snippet			

Return to Workload YAML reference table