Reserve training plans for your training jobs or HyperPod clusters

Amazon SageMaker training plans is a capability that allows you to reserve and help maximize the use of GPU capacity for large-scale AI model training workloads. This feature provides access to highly sought-after instance types that cover a range of GPU-accelerated computing options, including the latest NVIDIA GPU technologies and AWS trainium chips. With SageMaker training plans, you can secure predictable access to these high-demand, high-performance computational resources within your specified timelines and budgets, without the need to manage underlying infrastructure. This flexibility is particularly valuable for organizations dealing with the challenges of acquiring and scheduling these oversubscribed compute instances for their mission-critical AI workloads.

What are SageMaker training plans

SageMaker training plans allow you to reserve compute capacity tailored to your target resource needs, such as SageMaker training jobs or SageMaker HyperPod clusters. The service automatically handles the reservation, provisioning of accelerated compute resources, infrastructure setup, workload execution, and recovery from infrastructure failures.

SageMaker training plans consist of one or more Reserved Capacity blocks, each defined by the following parameters:

- Specific instance type
- Quantity of instances
- Availability Zone
- Duration
- Start and end times

Note

- Training plans are specific to their target resource (either SageMaker Training Job or SageMaker HyperPod) and cannot be interchanged.

- Multiple Reserved Capacity blocks in a single training plan may be discontinuous. This means there can be gaps between the Reserved Capacity blocks.

Benefits of SageMaker training plans

SageMaker training plans offer the following benefits:

- **Predictable Access**: Reserve GPU capacity for your machine learning workloads within specified time frames.
- **Cost Management**: Plan and budget for large-scale training requirements in advance.
- **Automated Resource Management**: SageMaker training plans handle the provisioning and management of infrastructure.
- **Flexibility**: Create training plans for various resources, including SageMaker training jobs and SageMaker HyperPod clusters.
- **Fault Tolerance**: Benefit from automatic recovery from infrastructure failures and workload migration across Availability Zones for SageMaker AI training jobs.

SageMaker training plans advance reservation and flexible start times

SageMaker training plans allow you to reserve compute capacity in advance, with flexible start times and durations.

- **Advance reservation**: You can reserve a training plan up to 8 weeks (56 days) in advance of the start date.
- **Minimum lead time**: SageMaker training plans offerings may be available to start within 30 minutes of reservation, subject to availability.

Note

You can search for and purchase a plan that will be accessible within 30 minutes. To ensure timely activation, the payment transaction must successfully complete at least 5 minutes before the desired start time. For example, if you want a plan to start at 2:00

PM, you can make a last-minute search as late as 1:30 PM and complete your purchase by 1:55 PM to guarantee the plan is ready by 2:00 PM.

- **Reservation duration and instance quantity**: SageMaker training plans allow you to reserve instances with specific duration and quantity options. For available instance types in a given AWS Region, duration, and quantity options, see the section called “Supported instance types, AWS Regions, and pricing”.
- **End time**: Training Plans always end at 11:30 AM UTC on the final day of the reservation.
- **Training plan termination**: If you're using training jobs as a target resource and 30 minutes remain in a Reserved Capacity, SageMaker training plans initiates the process of terminating any running instances within that block until the next Reserved Capacity becomes active. You retain full access to your training plan until 30 minutes before the final Reserved Capacity block's end time.

  If your target resource is a SageMaker HyperPod cluster, this time limit is one hour.

SageMaker training plans user workflow

SageMaker training plans work through the following steps:

Admin steps:

1. **Search and review**: Find available plan offerings that match your compute requirements, such as instance type, count, start time, and duration.
2. **Create a plan**: Reserve a training plan that meets your needs using the ID of your chosen plan offering.
3. **Payment and scheduling**: Upon successful upfront payment, the plan status becomes `Scheduled`.

Steps for plan users / ML engineers:

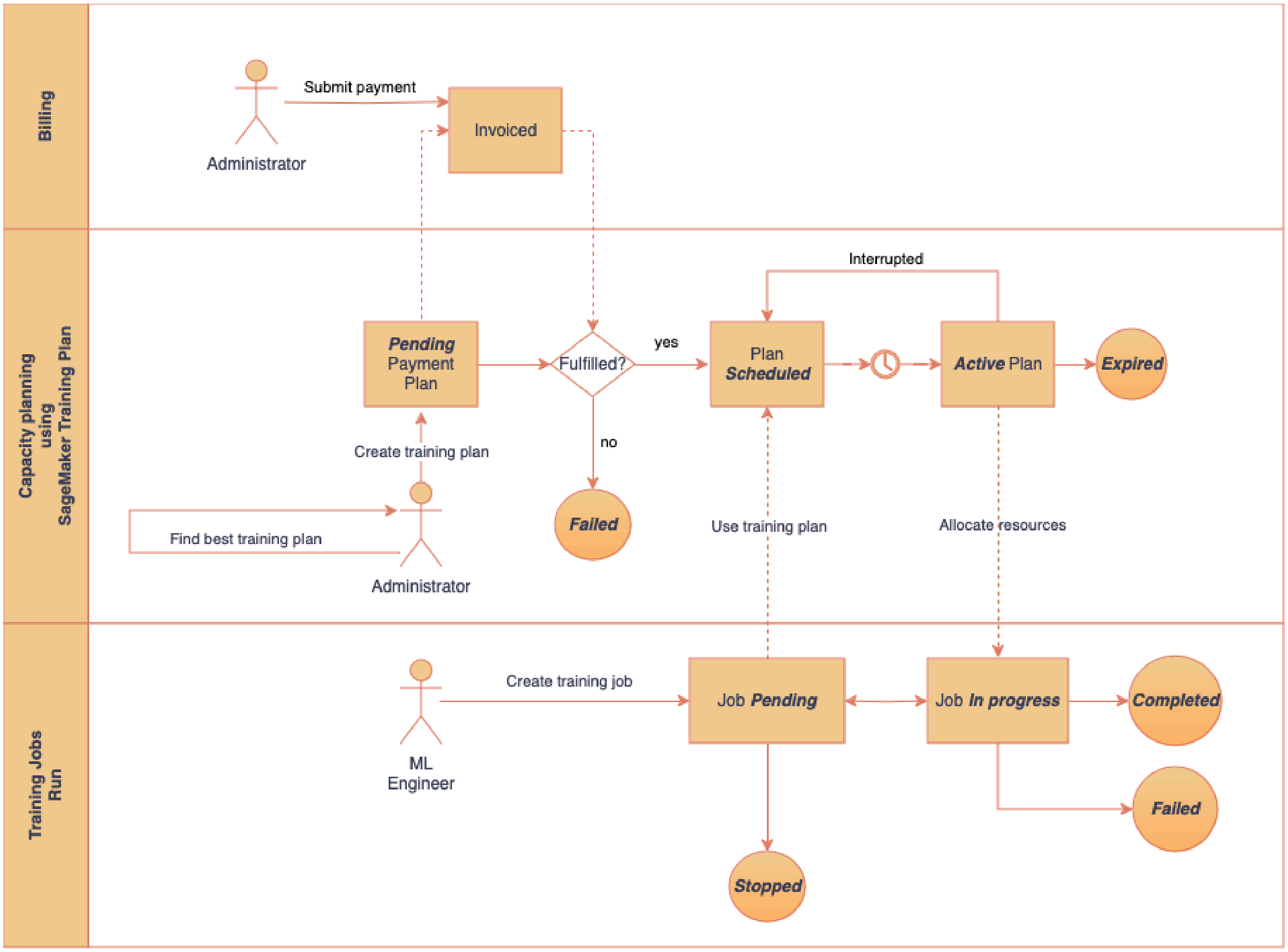
1. **Resource allocation**: Use your plan to queue SageMaker AI training jobs or allocate to a SageMaker HyperPod cluster instance group.
2. **Activation**: When the plan start date arrives, it becomes `Active`. Based on available reserved capacity, SageMaker training plans automatically launch training jobs or provision instance groups.

Note

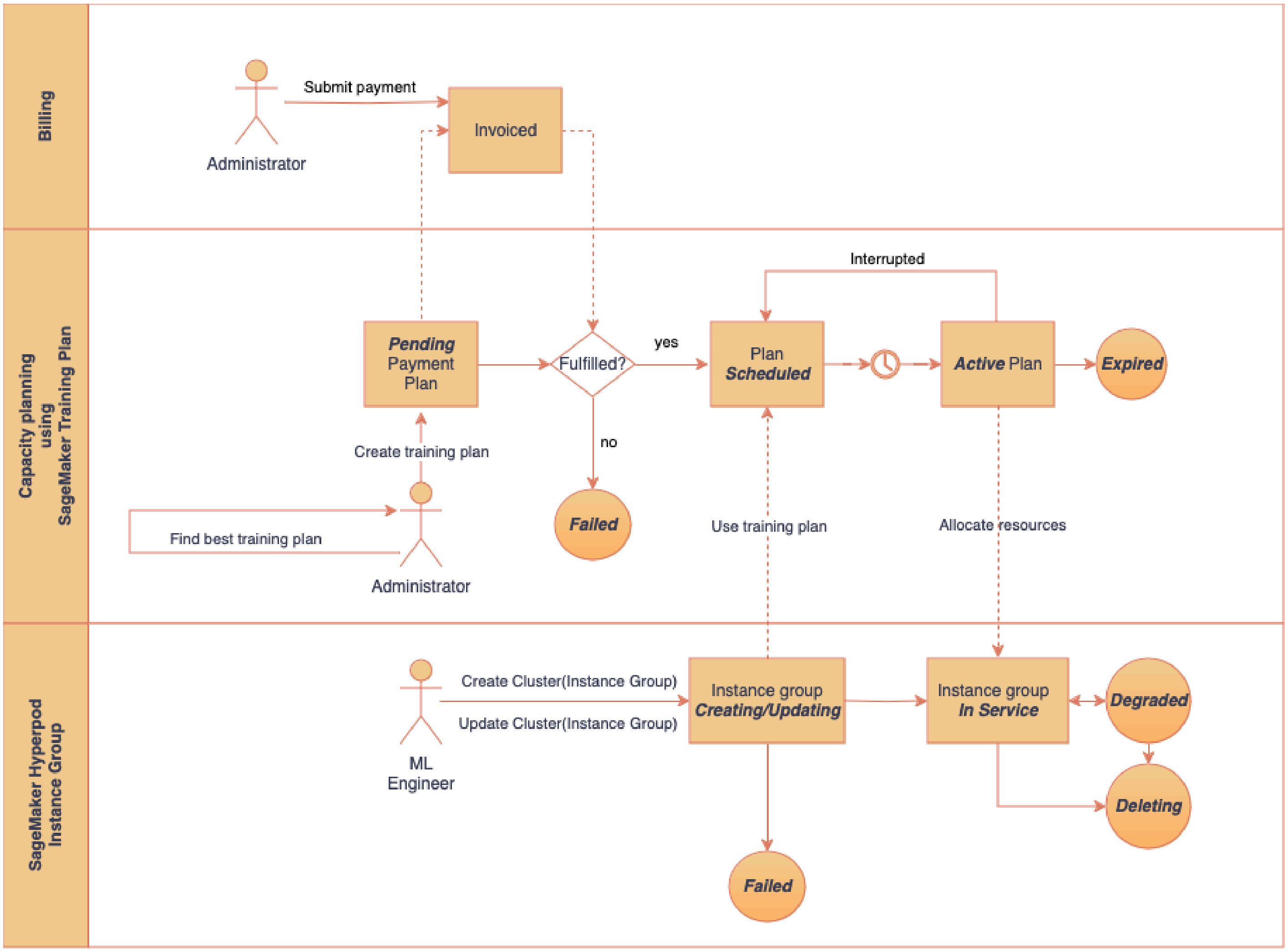
The status of the training plan transitions from `Scheduled` to `Active` when a Reserved Capacity period begins, and then back to `Scheduled` when waiting for the next Reserved Capacity period to start.

The following diagrams provide a comprehensive overview of how SageMaker training plans interact with different target resources, illustrating a plan's lifecycle and its role in resource allocation for both SageMaker training jobs and SageMaker HyperPod clusters.

- **Training plans for SageMaker Training Job**: The first diagram illustrates the end-to-end workflow of the interaction between a training plan and SageMaker Training Job.



- **Training plans for SageMaker HyperPod clusters**: The second diagram illustrates the end-to-end workflow of the interaction between a training plan and a SageMaker HyperPod instance group.



Supported instance types, AWS Regions, and pricing

Training plans support reservations for the following specific high-performance instance types, each available in select AWS Regions:

- **ml.p4d.24xlarge**
- **ml.p5.48xlarge**
- **ml.p5e.48xlarge**
- **ml.p5en.48xlarge**
- **ml.trn1.32xlarge**

- **ml.trn2.48xlarge**
- **ml.p6-b200.48xlarge**
- **ml.c6i-32xlargesc**

UltraServers

- **ml.p6e-gb200.36xlarge**
- **ml.p6e-gb200.72xlarge**

Note

The availability of instance types may change over time. For the most up-to-date information on available instance types according to Region, as well as their respective prices, see SageMaker Pricing. Scroll down to the **Amazon SageMaker HyperPod flexible training plans** section under **On-Demand Pricing**. Select a Region to view the list of available instance types.

The availability across multiple regions allows to choose the most suitable location for workloads, considering factors such as data residency requirements and proximity to other AWS services.

Important

- You can use SageMaker training plans to reserve instances with the following reservation duration and instance quantity options.
  - Reservation durations are available in 1-day increments from 1 to 182 days.
  - The reservation instance quantity options are 1, 2, 4, 8, 16, 32 or 64 instances.
- Make sure that your Training Jobs or HyperPod service quotas allow a maximum number of instances per instance type that exceeds the number of instances specified in your plan. To view your current quotas or request a quota increase, see the section called “Quotas and pricing”.

UltraServers in SageMaker AI

UltraServers in SageMaker AI offer a set of instances interconnected via a high bandwidth network domain. For example, the P6e-GB200 UltraServer connects up to 18 `p6e-gb200.36xlarge` instances under one NVIDIA NVLink domain. With 4 NVIDIA Blackwell GPUs per instance, each P6e-GB200 UltraServer supports 72 GPUs, so you can run your largest AI workloads with high performance on SageMaker AI.

When you use UltraServers with SageMaker AI, you get performance combined with SageMaker AI's managed infrastructure, built-in fault resiliency features, integrated monitoring capabilities, and native integration with other SageMaker AI and AWS services. This integration allows you to focus on model development and deployment while SageMaker AI handles the undifferentiated heavy lifting of managing AI infrastructure.

> **Note**
>
> UltraServers are available only in the Dallas Local Zone (us-east-1-dfw-2a), which is an extension of the US East (N. Virginia) Region. For more information, see Getting started with AWS Local Zones

Considerations

Consider the following when using UltraServers with SageMaker AI:

- You can use UltraServers for both SageMaker HyperPod and SageMaker training jobs.
- You can only purchase UltraServers in full units. For more information about instance and pricing information, see Amazon SageMaker HyperPod flexible training plans in Amazon SageMaker AI pricing.
- If you're using UltraServers with HyperPod, HyperPod automatically adds topology labels to your resources to help you with resource allocation. For more information, see Using topology-aware scheduling in Amazon SageMaker HyperPod.
- SageMaker AI and UltraServers offer various capabilities that enhance the resiliency of your workloads, including preemptive checks and automatic fault detection and mitigation. Depending on what the issue is, SageMaker AI can run actions to recover your workloads, such as restarting instances, replacing failed instances with spares, and replacing failed UltraServers.

- For added resilience, you can configure instances within an UltraServer to be used as spares. Keeping a spare instance within the UltraServer ensures that SageMaker AI can quickly respond to an instance failure while minimizing any impact to your jobs. We recommend that you keep one spare instance per UltraServer. You don't have to reserve any spare instances, but this might hinder support options and slow down failure recovery. You purchase UltraServers by wholes, so the number of spares that you reserve doesn't affect pricing.
- To see the status and instances within an UltraServer, use the ListTrainingPlans API operation or the AWS console to see training plans. Using these tools, you can see the total number of available instances, instances currently in use, unhealthy instances, the number of configured spares, and other information. Possible health statuses are `ok`, `impaired`, and `insufficient-data`.

SageMaker training plans search behavior

When searching for a training plan offering, SageMaker training plans use the following approach to maximize resource availability and flexibility for users, even when demand is high and Reserved Capacity blocks are scarce:

- **Initial continuous search**: SageMaker training plans first attempt to find a single, continuous block of Reserved Capacity that matches the specified duration within the start and end dates, while meeting all other specified criteria, including target resource, requested instance type, and number of instances.
- **Two-block search**: SageMaker training plans don't return a "no capacity" result if a single continuous Reserved Capacity block meeting all criteria is unavailable. Instead, it automatically attempts to fulfill the request using two separate Reserved Capacity blocks, splitting the total duration across two time segments.

  This two-block approach provides more flexibility in resource allocation, potentially securing high-demand instances that would otherwise be unavailable.

> **Note**
>
> SageMaker training plans return up to three offerings of one or two segments. For example, for a 48-hour duration plan, SageMaker training plans might offer a plan with two 24-hour blocks, one continuous 48-hour block, and two blocks with uneven duration.

Considerations

> ⚠ **Important**
>
> - Training plans cannot be modified once purchased.
> - Training plans cannot be shared across AWS accounts or within your AWS Organization.

- When searching for training plan offerings, SageMaker training plans adapts its search strategy based on the target resources:

  For SageMaker HyperPod clusters:

  - Offerings are limited to a single Availability Zone (AZ).
  - This ensures consistent network performance and data locality within the cluster.

  For SageMaker training jobs:

  - Offerings can span multiple Availability Zones.
  - This is particularly relevant when the plan offering contains multiple discontinuous reserved capacities.
  - For example, a plan might include capacity in AZ-A for one Reserved Capacity block and AZ-B for another. SageMaker training plans can automatically move workloads across Availability Zones (AZs) based on resource availability.

    This multi-AZ approach for training jobs provides greater flexibility in resource allocation, increasing the chances of finding suitable capacity for your workload. However, you should be aware that your jobs may run in different AZs during different parts of your reservation period.
- When presented with a two-block offering, users should carefully consider if this split allocation meets their workload requirements. This may require adjusting job scheduling or workload distribution to accommodate the non-continuous nature of the reservation.

IAM for SageMaker training plans

SageMaker training plans requires specific permissions for two distinct roles::

1. **Plan creator role**: Users assigned the *Plan Creator* role need permissions to search training plan offerings, create new training plans, list and describe training plans.

2. **Plan user role**: Users with the *Plan User* role require permissions to use training plans in SageMaker training jobs or when creating and updating SageMaker HyperPod clusters.

Before using SageMaker training plans, update permissions based on your access method:

- For AWS Management Console or SageMaker SDKs users: Update the permissions of the IAM role configured for the console user or API user.
- For AWS CLI users: Ensure your AWS CLI profile is correctly configured with the appropriate credentials and permissions.
- For Studio application users such as JupyterLab, set permissions on the execution role associated with the space used by the application.

You can set these permissions using either a managed policy or individual more granular permissions.

For information about how to update the permissions policy for a role, see Update permissions for a role. For information about how to find and update an execution role, see Get your execution role.

> **Note**
>
> Administrators should carefully consider which users need the ability to create training plans and assign permissions accordingly.

Managed policies

- For plan creators: `AmazonSageMakerTrainingPlanCreateAccess` provides access to create and manage training plans.
- For plan users: `AmazonSageMakerFullAccess` includes the permissions to use training plans.

> **Note**
>
> - The `AmazonSageMakerFullAccess` managed policy is designed as an ease-of-use policy primarily for experimentation purposes. While it provides broad access to SageMaker AI features, including the use of training plans, it's important to note:

- This policy is not recommended for production environments due to its broad permissions.
- It does not include permissions for creating training plans, as `CreateTrainingPlan` is considered an administrative action requiring upfront payment.
- For production use cases, we strongly recommend creating custom policies that adhere to the principle of least privilege, granting only the specific permissions required for each role.

Individual permissions

The following list details the granular permissions that should be set in the IAM policy statements of a role, based on the specific actions a user needs to perform with SageMaker training plans:

Training plans list of permissions

- `SearchTrainingPlanOfferings`: This permission allows users to search for available training plan offerings.

```
{
  "Sid": "SearchTrainingPlanOfferingsPermissions",
  "Effect": "Allow",
  "Action": [
    "sagemaker:SearchTrainingPlanOfferings"
  ],
  "Resource": "*"
}
```

- `CreateTrainingPlan`: This permission allows users to create new training plans.

Note

You must also include permissions for `CreateReservedCapacity` and `AddTags`, and specify both `training-plan` and `reserved-capacity` resource types.

```
{
  "Sid": "CreateTrainingPlanPermissions",
  "Effect": "Allow",
  "Action": [
    "sagemaker:CreateTrainingPlan",
    "sagemaker:CreateReservedCapacity",
    "sagemaker:AddTags"
  ],
  "Resource": [
    "arn:aws:sagemaker:*:*:training-plan/*",
    "arn:aws:sagemaker:*:*:reserved-capacity/*"
  ]
}
```

- `DescribeTrainingPlan` : This permission allows users to view details of existing training plans.

```
{
  "Sid": "DescribeTrainingPlanPermissions",
  "Effect": "Allow",
  "Action": [
    "sagemaker:DescribeTrainingPlan"
  ],
  "Resource": [
    "arn:aws:sagemaker:::training-plan/*"
  ]
}
```

- `ListTrainingPlans`: This permission allows users to list all training plans in their AWS account.

```
{
  "Sid": "ListTrainingPlansPermissions",
  "Effect": "Allow",
  "Action": [
    "sagemaker:ListTrainingPlans"
  ],
  "Resource": "*"
}
```

Individual permissions per type of user

This section provides a detailed breakdown of the individual permissions required for each role, as mentioned in the the section called "IAM for SageMaker training plans" section.

For plan creators, the following permissions are necessary:

- `sagemaker:SearchTrainingPlanOfferings`
- `sagemaker:CreateTrainingPlan`
- `sagemaker:CreateReservedCapacity`
- `sagemaker:AddTags`
- `sagemaker:DescribeTrainingPlan`
- `sagemaker:ListTrainingPlans`

Plan users require these permissions:

- `sagemaker:CreateTrainingJob` (for SageMaker Training Job)
- `sagemaker:CreateCluster` and `sagemaker:UpdateCluster` (for SageMaker HyperPod)
- Access to the `training-plan` and `reserved-capacity` resources; When configuring IAM policies for SageMaker training plans, include permissions for both `training-plan` and `reserved-capacity` resources. These resources are required for both SageMaker training jobs and SageMaker HyperPod clusters. This allows your IAM roles to interact with SageMaker training plans resources and manage Reserved Capacity.
  - For SageMaker training jobs, ensure your policy includes the `"arn:aws:sagemaker:::training-plan/"` and `"arn:aws:sagemaker:::reserved-capacity/"` resource ARNs.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "sagemaker:CreateTrainingJob"
      ],
      "Resource": [
        "arn:aws:sagemaker:us-east-2:111122223333:training-job/",
        "arn:aws:sagemaker:us-east-2:111122223333:training-plan/",
        "arn:aws:sagemaker:us-east-2:111122223333:reserved-capacity/*"
        ]
      }
    ]
  }
```

Similarly, for SageMaker HyperPod configurations, include these same ARNs in addition to the cluster-specific resources.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "sagemaker:CreateCluster",
        "sagemaker:UpdateCluster"
      ],
      "Resource": [
        "arn:aws:sagemaker:us-east-2:111122223333:cluster/",
        "arn:aws:sagemaker:us-east-2:111122223333:training-plan/",
        "arn:aws:sagemaker:us-east-2:111122223333:reserved-capacity/*"
      ]
    }
  ]
}
```

Training plans creation

To reserve compute capacity using the SageMaker training plans capability, follow these steps:

1. **Identify your target resource:** Begin by determining whether you need capacity for SageMaker training jobs or SageMaker HyperPod clusters.
2. **Specify your capacity requirements :** Define your capacity needs in detail. This includes selecting the appropriate instance type for your workload, determining the number of instances required, and specifying the duration of use. For information about the supported instance

types in a given AWS Region, duration, and quantity options, see the section called "Supported instance types, AWS Regions, and pricing".

3. **Search for available training plan offerings:** Once you specify your requirements, use SageMaker training plans' search functionality to find available training plan offerings across one or more segments. Each offering includes details such as start time, specific availability zone for Reserved Capacity, and plan price. Review these offerings, considering factors like cost-effectiveness, geographical preferences, and alignment with your specified needs.

   If no suitable plan is available, adjust your search criteria and look for a new set of offerings.

4. **Create a training plan based on a suitable offering:** After identifying a suitable offering, proceed to create your training plan. This process involves selecting your chosen offering and initiating the reservation.

   - The training plan reservation creates an invoice.
   - The payment for the total amount is collected as part of the fulfillment process. Once the payment is completed, the plan is ready for scheduling your SageMaker training jobs or creating HyperPod clusters.

   To learn about how to use training plans for your SageMaker training jobs , see the section called "Training plans utilization for SageMaker training jobs".

   To learn about how to use training plans for your HyperPod clusters, see the section called "Training plans utilization for SageMaker HyperPod clusters".

You can create a training plan using either the SageMaker AI console or programmatic methods. The SageMaker AI console offers a visual, graphical interface with a comprehensive view of your options, while programmatic creation can be done using the AWS CLI or SageMaker SDKs to interact directly with the training plans API.

For step-by-step console instructions and detailed API references, refer to the respective sections in this documentation.

Topics

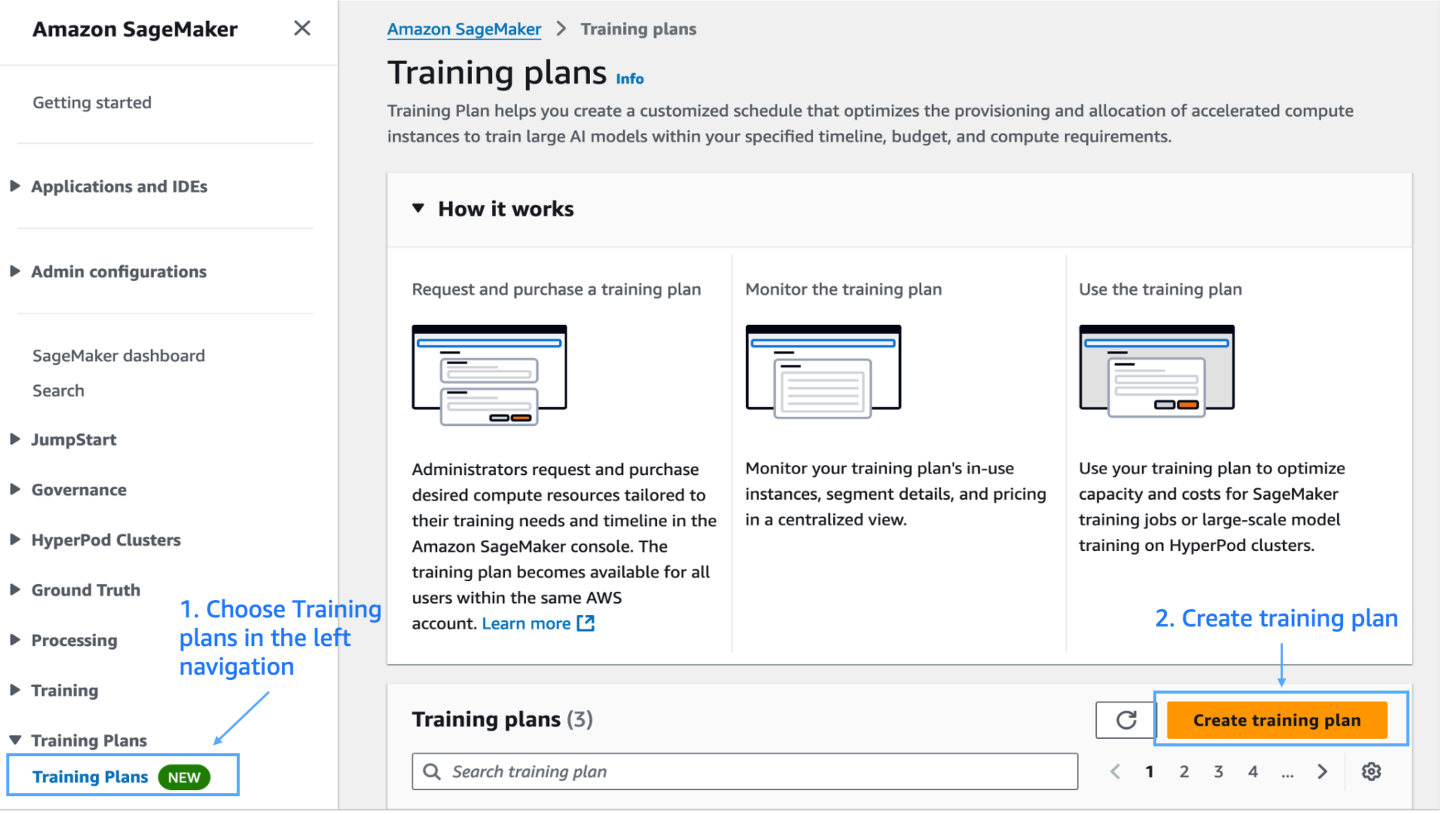
- SageMaker training plans creation using the SageMaker AI console
- SageMaker training plans creation using the SageMaker API, or AWS CLI

SageMaker training plans creation using the SageMaker AI console

SageMaker training plans offer a convenient way to create training plans through the SageMaker AI console UI, allowing users to easily schedule their machine learning training resources. This guide walks you through the process of creating a training plan for SageMaker training jobs and SageMaker HyperPod clusters using the SageMaker AI console. By following these steps, you will search for training plan offerings, review available options, and purchase the plan that best fits your needs.

To create a training plan visually using a UI:

1. Start by navigating to the SageMaker AI console at https://console.aws.amazon.com/sagemaker/.
2. Choose **Training Plans** in the left pane menu.
3. From there, choose the **Create training plan** button in the main content area to start the process of setting up your customized training schedule.



Training plans Info

Training Plan helps you create a customized schedule that optimizes the provisioning and allocation of accelerated compute instances to train large AI models within your specified timeline, budget, and compute requirements.

How it works

Request and purchase a training plan

Administrators request and purchase desired compute resources tailored to their training needs and timeline in the Amazon SageMaker console. The training plan becomes available for all users within the same AWS account. Learn more

Monitor the training plan

Monitor your training plan's in-use instances, segment details, and pricing in a centralized view.

Use the training plan

Use your training plan to optimize capacity and costs for SageMaker training jobs or large-scale model training on HyperPod clusters.

1. Choose Training plans in the left navigation

2. Create training plan

Next, search for plan offerings that match your compute requirements.

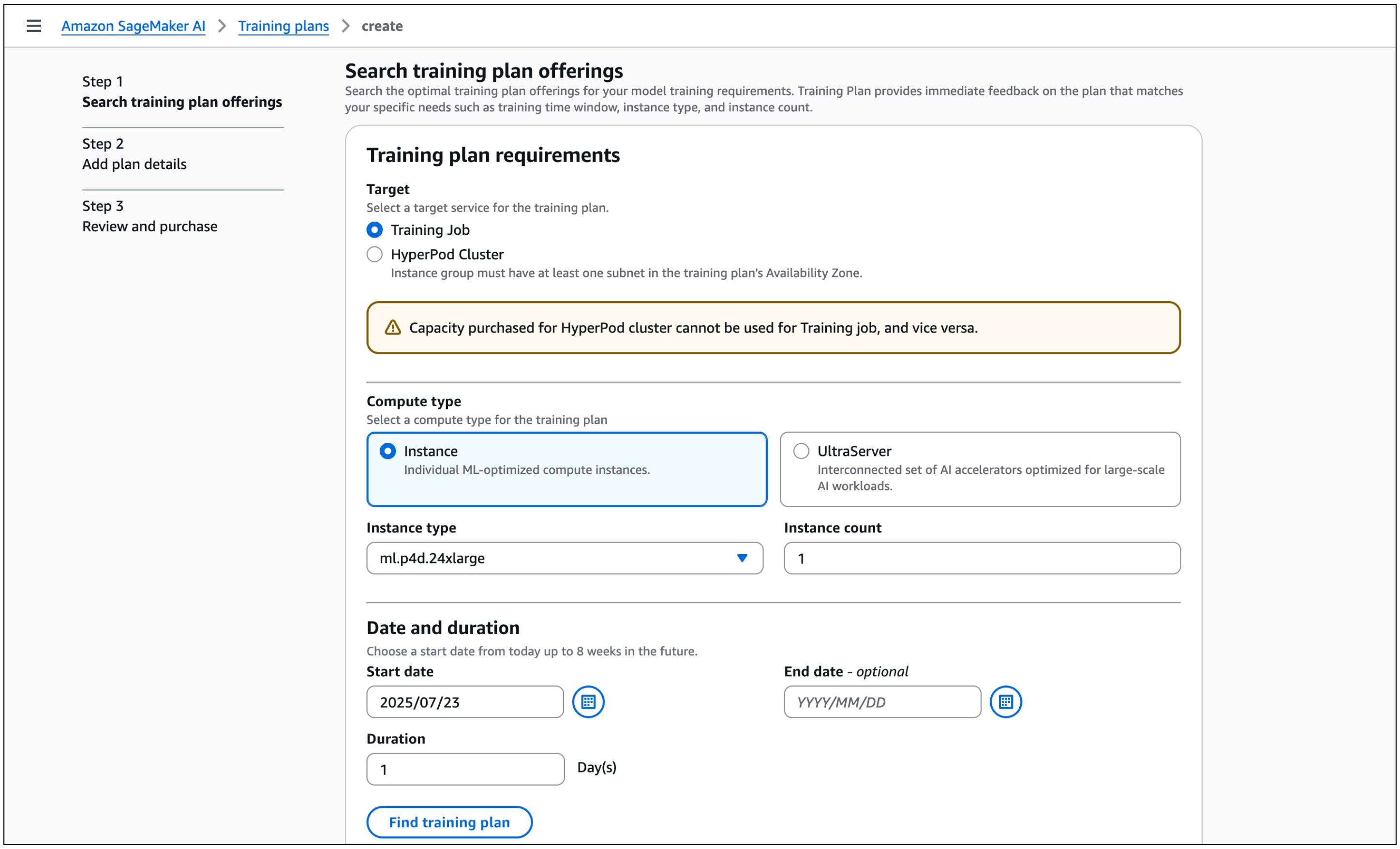
Topics

Search training plan offerings

After you choose **Training Plans** in the left pane of the SageMaker AI console, and then **Create training plan**, a **Find training plan** form opens up. This form allows you to specify your requirements and search for suitable training plan offerings.

Follow these steps to complete the form:

1. Identify your **Target**: Training plans are specific to their target resource. Specify whether you want to use a plan to run SageMaker training jobs or SageMaker HyperPod clusters.
2. For **Compute type**, you can choose between **Instance** or **UltraServer**. UltraServers are connect multiple Amazon EC2 instances using a low-latency, high-bandwidth accelerator interconnect. For more information, see Amazon EC2 UltraServers. To learn about how you can use UltraServers with SageMaker AI, see UltraServers in SageMaker AI.
3. Choose your preferred **Instance type** and **Instance count**: For available instance types in a given AWS Region, duration, and quantity options, see the section called "Supported instance types, AWS Regions, and pricing".
4. Define your time parameters: Choose your desired start and end dates, and specify the plan duration within this window.
5. Choose **Find training plans**.



Amazon SageMaker AI > Training plans > create

Step 1
Search training plan offerings

Step 2
Add plan details

Step 3
Review and purchase

Search training plan offerings

Search the optimal training plan offerings for your model training requirements. Training Plan provides immediate feedback on the plan that matches your specific needs such as training time window, instance type, and instance count.

Training plan requirements

Target

Select a target service for the training plan.

● Training Job

○ HyperPod Cluster
Instance group must have at least one subnet in the training plan's Availability Zone.

⚠ Capacity purchased for HyperPod cluster cannot be used for Training job, and vice versa.

Compute type

Select a compute type for the training plan

● Instance
Individual ML-optimized compute instances.

○ UltraServer
Interconnected set of AI accelerators optimized for large-scale AI workloads.

Instance type: ml.p4d.24xlarge

Instance count: 1

Date and duration

Choose a start date from today up to 8 weeks in the future.

Start date: 2025/07/23

End date - *optional*: YYYY/MM/DD

Duration: 1 Day(s)

Find training plan

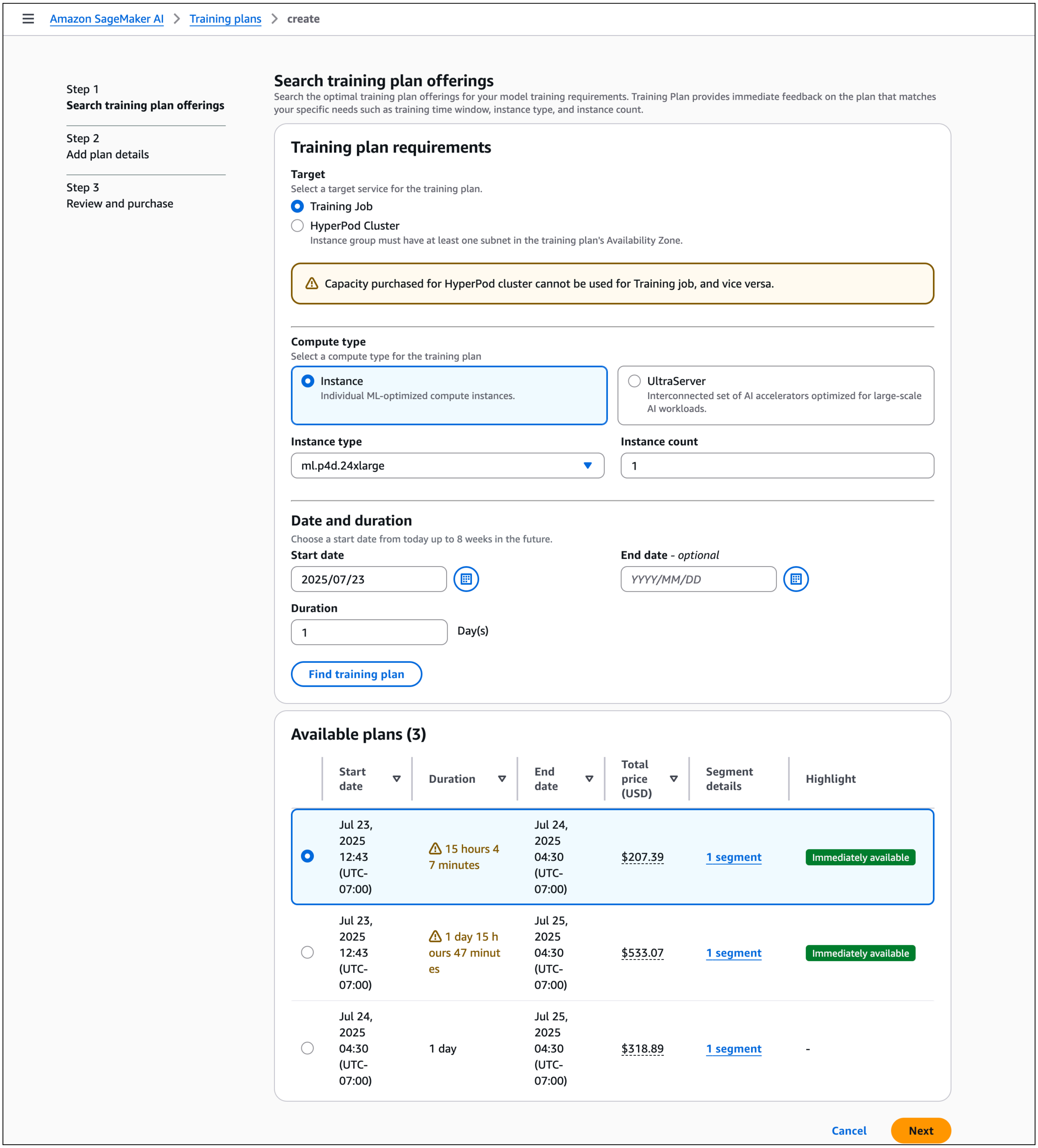
SageMaker training plans search for offerings that match your capacity requirements. When matches are found within your specified time frame, they appear at the bottom of the page. Each training plan offering includes the following details:

- Total plan duration
- Start and end dates
- Total upfront price:

  Hover over the price to view the detailed breakdown of instance hourly rate, instance count, and total hours
- Total number of plan segments

Clicking the segment detail link opens a modal view with segment-specific details:

- Duration
- Start and end dates
- Availability zone



Step 1
Search training plan offerings

Step 2
Add plan details

Step 3
Review and purchase

Search training plan offerings

Search the optimal training plan offerings for your model training requirements. Training Plan provides immediate feedback on the plan that matches your specific needs such as training time window, instance type, and instance count.

Training plan requirements

Target
Select a target service for the training plan.

● Training Job
○ HyperPod Cluster
Instance group must have at least one subnet in the training plan's Availability Zone.

⚠ Capacity purchased for HyperPod cluster cannot be used for Training job, and vice versa.

Compute type
Select a compute type for the training plan

● Instance
Individual ML-optimized compute instances.

○ UltraServer
Interconnected set of AI accelerators optimized for large-scale AI workloads.

Instance type: ml.p4d.24xlarge

Instance count: 1

Date and duration

Choose a start date from today up to 8 weeks in the future.

Start date: 2025/07/23

End date - *optional*: YYYY/MM/DD

Duration: 1 Day(s)

Available plans (3)

	Start date	Duration	End date	Total price (USD)	Segment details	Highlight
●	Jul 23, 2025 12:43 (UTC-07:00)	⚠ 15 hours 47 minutes	Jul 24, 2025 04:30 (UTC-07:00)	$207.39	1 segment	Immediately available
○	Jul 23, 2025 12:43 (UTC-07:00)	⚠ 1 day 15 hours 47 minutes	Jul 25, 2025 04:30 (UTC-07:00)	$533.07	1 segment	Immediately available
○	Jul 24, 2025 04:30 (UTC-07:00)	1 day	Jul 25, 2025 04:30 (UTC-07:00)	$318.89	1 segment	-

If no suitable plans are found or the available plans don't meet your needs, adjust your search criteria by modifying the parameters in the **Training plans requirements** form. Once you find a suitable offering, select it and choose **Next** to continue to the plan reservation page. On this page, you can name your plan, and then review and confirm your selection before finalizing your reservation.

> **Note**
>
> Plans marked `Immediately available` will start within 30 minutes, provided payment is completed no less than 5 minutes before the scheduled start time.

Reserve the best training plan

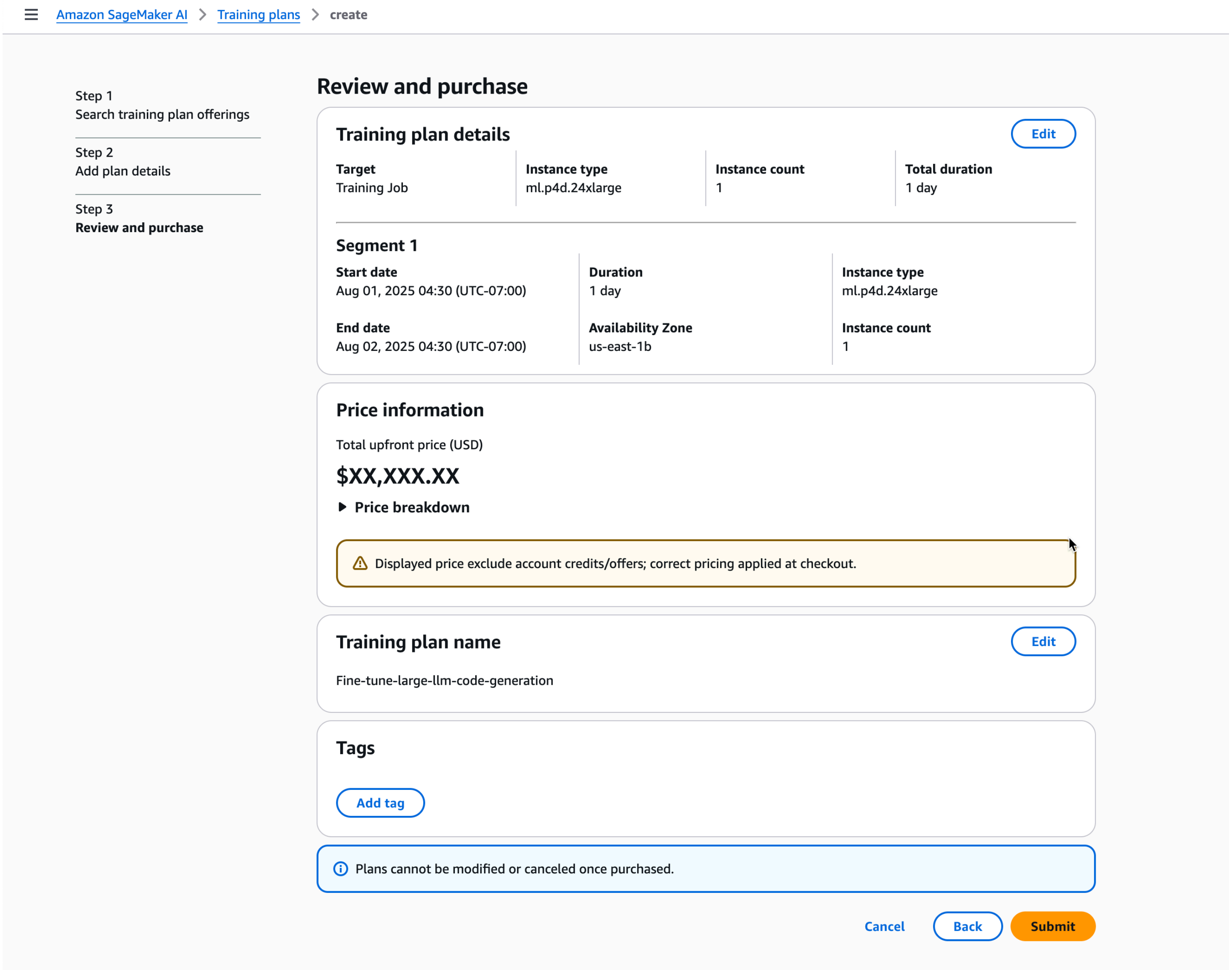
The search of a training plan has returned offerings that fit your capacity needs and budget.

1. Enter a name for your plan and then choose **Next**.
2. Review and **Submit** your purchase order.

   > **Important**
   >
   > - Training plans cannot be modified once purchased.
   > - Training plans cannot be shared across AWS accounts or within your AWS Organization.

   After submitting your order

   - The training plan initially appears as `Pending` in your training plan list.
   - An invoice is generated automatically upon order receipt.
   - The total payment is collected during the fulfillment process.
   - Once payment is successfully processed, the plan status changes to `Scheduled` and the plan becomes available for use.



Step 1
Search training plan offerings

Step 2
Add plan details

Step 3
Review and purchase

Review and purchase

Training plan details

Target
Training Job

Instance type
ml.p4d.24xlarge

Instance count
1

Total duration
1 day

Segment 1

Start date
Aug 01, 2025 04:30 (UTC-07:00)

Duration
1 day

Instance type
ml.p4d.24xlarge

End date
Aug 02, 2025 04:30 (UTC-07:00)

Availability Zone
us-east-1b

Instance count
1

Price information

Total upfront price (USD)

$XX,XXX.XX

▶ **Price breakdown**

⚠ Displayed price exclude account credits/offers; correct pricing applied at checkout.

Training plan name

Fine-tune-large-llm-code-generation

Tags

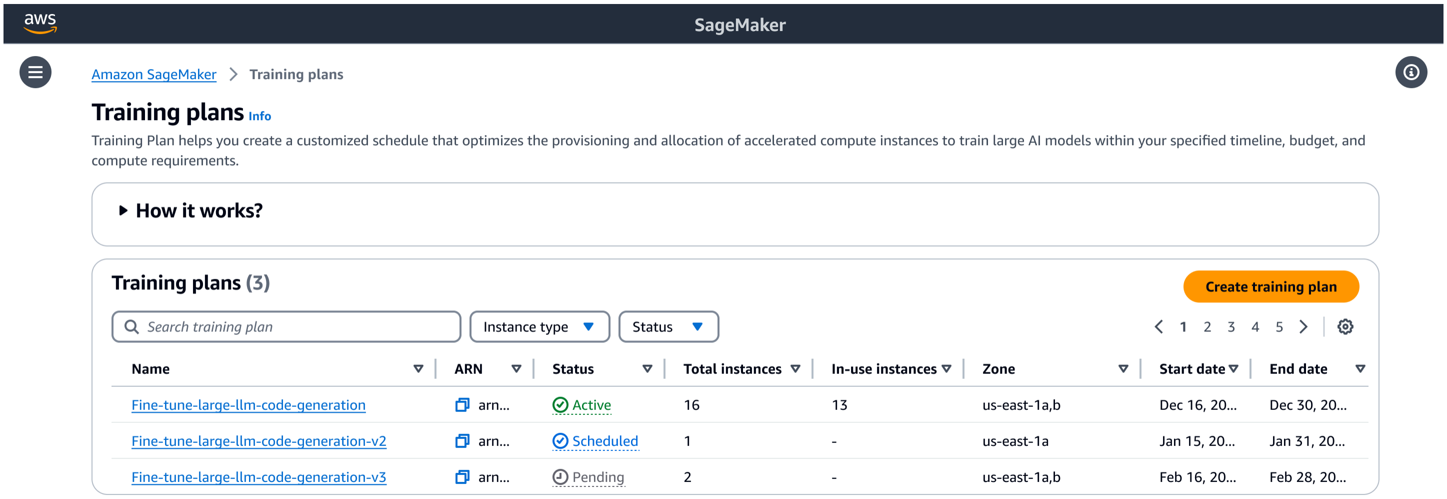
ⓘ Plans cannot be modified or canceled once purchased.

List training plans

To view your training plans:

1. Navigate to the SageMaker AI console at https://console.aws.amazon.com/sagemaker/.
2. Choose **Training Plans** in the left pane menu. This displays a list of all your training plans, including their names, status, target resource type, and other key details.

   After purchasing a plan, you are directed to this list. Newly created plans appear with a `Pending` status until payment is completed. The status is typically updated within a few minutes of payment processing.



Training plans Info

Training Plan helps you create a customized schedule that optimizes the provisioning and allocation of accelerated compute instances to train large AI models within your specified timeline, budget, and compute requirements.

Name	ARN	Status	Total instances	In-use instances	Zone	Start date	End date
Fine-tune-large-llm-code-generation	arn...	Active	16	13	us-east-1a,b	Dec 16, 20...	Dec 30, 20...
Fine-tune-large-llm-code-generation-v2	arn...	Scheduled	1	-	us-east-1a	Jan 15, 20...	Jan 31, 20...
Fine-tune-large-llm-code-generation-v3	arn...	Pending	2	-	us-east-1a,b	Feb 16, 20...	Feb 28, 20...

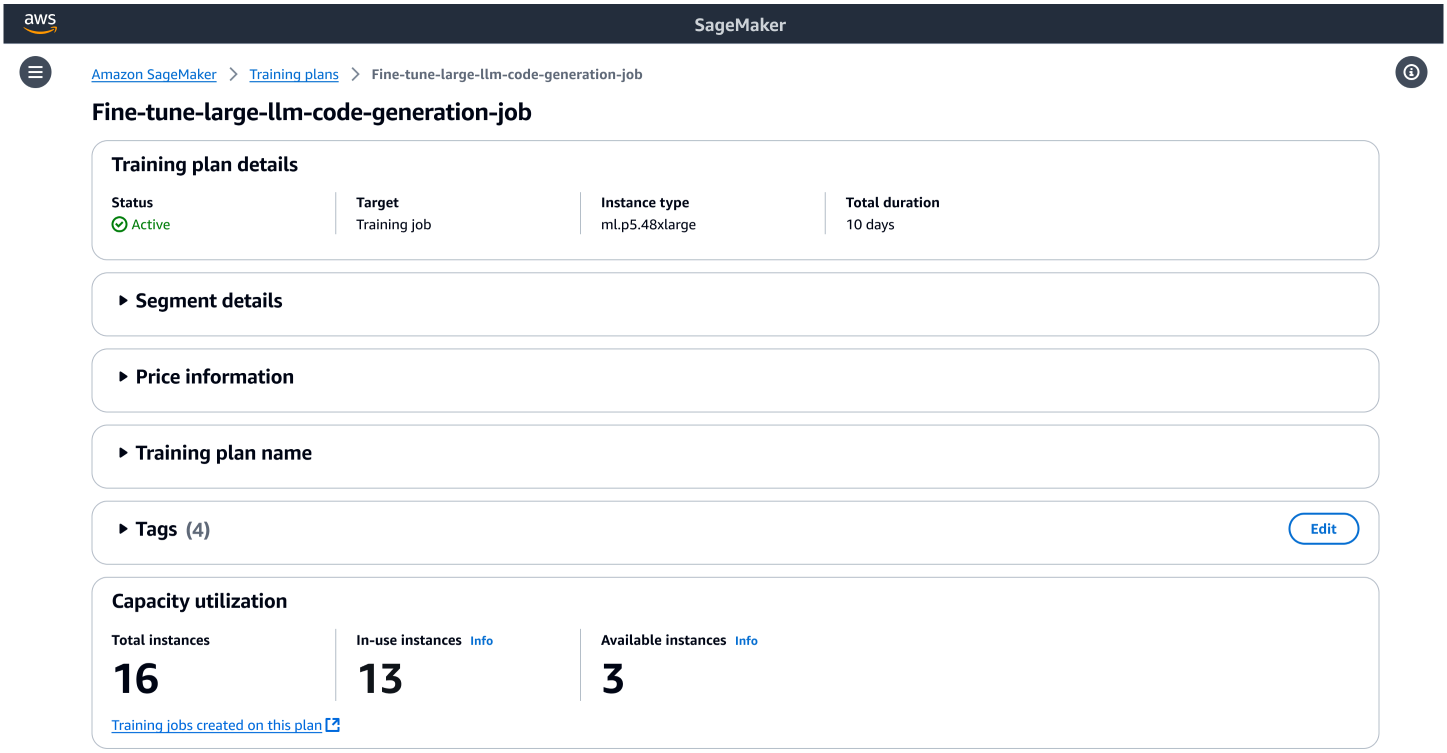
View training plan details

From the training plans list, follow a plan's name to view its details. Specifically, you can check your current capacity usage, and list your workloads in your plan's details page.

The details page shows:

- The training plan overview: Status, target, instance type, and duration.
- Expandable sections for segment details, pricing, plan name, and tags.
- Capacity utilization:
  - Total: The total number of instances reserved in this training plan.
  - In-use: The number of instances currently in use from this training plan.
  - Available instances: The number of instances currently available for use in this training plan.

At the bottom of the page, a link allows you to view either the training jobs or the list of SageMaker HyperPod cluster instance groups associated with this plan, depending on its target resource.



aws
SageMaker
Amazon SageMaker > Training plans > Fine-tune-large-llm-code-generation-job

Fine-tune-large-llm-code-generation-job

Training plan details

Status	Target	Instance type	Total duration
✓ Active	Training job	ml.p5.48xlarge	10 days

▸ Segment details

▸ Price information

▸ Training plan name

▸ Tags (4) Edit

Capacity utilization

Total instances	In-use instances Info	Available instances Info
16	13	3

Training jobs created on this plan

SageMaker training plans creation using the SageMaker API, or AWS CLI

SageMaker training plans support the programmatic creation of training plans through its API. You can interact with the training plans API using the AWS CLI or SageMaker SDKs.

SageMaker training plans's API actions provide a comprehensive workflow for managing training plans programmatically:

- **`SearchTrainingPlanOfferings`:** Enables users to query and discover available compute resources by specifying parameters like instance type, count, and desired time window. The API returns a ranked list of training plan offerings that best match the user's requirements.
- **`CreateTrainingPlan`:** Allows reservation of a specific training plan offering, transforming a potential compute capacity into scheduled reserved capacities with a unique training plan ARN.
- **`ListTrainingPlans`:** Provides a method to retrieve and review all existing training plans in a user's AWS account, with optional filtering and sorting capabilities.
- **`DescribeTrainingPlan`:** Offers detailed insights into a specific training plan, including its lifecycle stages from `Pending` to `Active` to `Expired`.

Topics

- Search training plan offerings
- Reserve the best training plan
- List training plans
- View training plan details

Search training plan offerings

To create a training plan, start by calling the `SearchTrainingPlanOfferings` API operation, passing your plan requirements (such as instance type, count, and desired time window) as input parameters. Training plans are specific to their target resource. Ensure that you specify which target resource the plan will be used for (`training-job` or `hyperpod-cluster`). The API returns a list of available offerings that match your requirements. If no suitable offerings are found, you may need to adjust your requirements and search again.

This API call retrieves the training plan offerings that best meet your capacity needs. Each `TrainingPlanOffering` returned in the response is identified by a unique offering ID. The first offering in the list represents the best match for your requirements. If no suitable training plan is available within your specified dates, the list is empty. Adjust your search criteria and look for a new set of offerings.

- Reservation durations are available in 1-day increments from 1 to 182 days.
- The reservation instance quantity options are 1, 2, 4, 8, 16, 32 or 64 instances.

To learn about the list of available instances supported by SageMaker training plans, see Supported instance types, AWS Regions, and pricing.

The following example uses an AWS CLI command to request training plan offerings with a specified instance type, count, and time information.

```
# List training plan offerings with instance type, instance count, duration in hours,
 start time after, and end time before.
aws sagemaker search-training-plan-offerings \
--target-resources "training-job" \
--instance-type "ml.p4d.24xlarge" \
--instance-count 1 \
--duration-hours 15 \
--start-time-after "1737484800"
```

```
--end-time-before "1737657600"
```

This JSON document is a sample response from the SageMaker training plans API. The response provides information about multiple available training plan offerings that match the specified capacity requirements. It includes three distinct offerings with varying durations, upfront fees, and start/end times, all using the same instance type and targeting training jobs.

```
{
    "TrainingPlanOfferings": [
        {
            "TrainingPlanOfferingId": "tpo-SHA-256-hash-value",
            "TargetResources": [
                "training-job"
            ],
            "RequestedStartTimeAfter": "2025-01-21T11:08:27.704000-08:00",
            "DurationHours": 15,
            "DurationMinutes": 51,
            "UpfrontFee": "xxxx.xx",
            "CurrencyCode": "USD",
            "ReservedCapacityOfferings": [
                {
                    "InstanceType": "ml.p4d.24xlarge",
                    "InstanceCount": 1,
                    "AvailabilityZone": "us-west-2a",
                    "DurationHours": 15,
                    "DurationMinutes": 51,
                    "StartTime": "2025-01-21T11:39:00-08:00",
                    "EndTime": "2025-01-22T03:30:00-08:00"
                }
            ]
        },
        {
            "TrainingPlanOfferingId": "tpo-SHA-256-hash-value",
            "TargetResources": [
                "training-job"
            ],
            "RequestedStartTimeAfter": "2025-01-21T11:08:27.704000-08:00",
            "DurationHours": 39,
            "DurationMinutes": 51,
            "UpfrontFee": "xxxx.xx",
            "CurrencyCode": "USD",
            "ReservedCapacityOfferings": [
                {
                    "InstanceType": "ml.p4d.24xlarge",
                    "InstanceCount": 1,
                    "AvailabilityZone": "us-west-2a",
                    "DurationHours": 39,
                    "DurationMinutes": 51,
                    "StartTime": "2025-01-21T11:39:00-08:00",
                    "EndTime": "2025-01-23T03:30:00-08:00"
                }
            ]
        },
        {
            "TrainingPlanOfferingId": "tpo-SHA-256-hash-value",
            "TargetResources": [
                "training-job"
            ],
            "RequestedStartTimeAfter": "2025-01-21T11:08:27.704000-08:00",
            "DurationHours": 24,
            "DurationMinutes": 0,
            "UpfrontFee": "xxxx.xx",
            "CurrencyCode": "USD",
            "ReservedCapacityOfferings": [
                {
                    "InstanceType": "ml.p4d.24xlarge",
                    "InstanceCount": 1,
                    "AvailabilityZone": "us-west-2a",
                    "DurationHours": 24,
                    "DurationMinutes": 0,
                    "StartTime": "2025-01-22T03:30:00-08:00",
                    "EndTime": "2025-01-23T03:30:00-08:00"
                }
            ]
        }
    ]
}
```

The following is a sample command of how to use the AWS CLI to search for training plan offerings that include UltraServers.

```
aws sagemaker search-training-plan-offerings \
--ultra-server-type ml.c6i-32xlargesc \
--ultra-server-count 1 \
--duration-hours 24 \
--target-resources hyperpod-cluster
```

```
--start-time-after "1737484800" \
--end-time-before "1737657600"
```

```
{
    "TrainingPlanOfferings": [
        {
            "TrainingPlanOfferingId": "tpo-SHA-256-hash-value",
            "TargetResources": [
                "training-job"
            ],
            "RequestedStartTimeAfter": "2025-07-21T16:59:25.760000+00:00",
            "DurationHours": 24,
            "DurationMinutes": 0,
            "UpfrontFee": "0.24",
            "CurrencyCode": "USD",
            "ReservedCapacityOfferings": [
                {
                    "ReservedCapacityType": "UltraServer",
                    "UltraServerType": "ml.u-p6e-gb200x72",
                    "UltraServerCount": 1,
                    "InstanceType": "ml.p6e-gb200.36xlarge",
                    "InstanceCount": 18,
                    "AvailabilityZone": "us-east-2a",
                    "DurationHours": 24,
                    "DurationMinutes": 0,
                    "StartTime": "2025-07-22T11:30:00+00:00",
                    "EndTime": "2025-07-23T11:30:00+00:00"
                }
            ]
        }
    ]
}
```

The following sections define the mandatory and optional input request parameters for the `SearchTrainingPlanOfferings` API operation.

Required parameters

When calling the `SearchTrainingPlanOfferings` API to list training plan offerings that meet your requirements, you must provide the following values:

- `TargetResources`: The target resources (`training-job` or `hyperpod-cluster`) for which the plan will be used. The default value is `training-job`. Training plans are specific to their target resource.
  - A training plan designed for SageMaker training jobs can only be used to schedule and run training jobs.
  - A training plan for HyperPod clusters can be used exclusively to provide compute resources to a cluster's instance group.
- `InstanceType`: The type of instance to provision. The `InstanceType` must be of a supported type.

  To learn about the list of available instances supported by SageMaker training plans, see [Supported instance types, AWS Regions, and pricing](#).
- `InstanceCount`: The number of instances to provision. If the number of instances is greater than 1, it should be a power of 2.
- `DurationHour`: The total duration of your requested plan in hours. The `DurationHour` is rounded up to the nearest multiple of 24.

Optional parameters

The following sections provide details on some optional parameters that you can pass to your `SearchTrainingPlanOfferings` API request.

- `StartTimeAfter`: Specify the requested start time of the plan. The `StartTimeAfter` should be a `timestamp` or an `ISO 8601 date/time` value in the future.
- `EndTimeBefore`: Specify the requested end time of the plan in a `timestamp` or an `ISO 8601 date/time` format. The `EndTimeBefore` should be at least 24 hours after the start time .
- `UltraServerType` : Specify the type of UltraServer to search for. For more information about UltraServers, see [UltraServers in SageMaker AI](#).
- `UltraServerCount`: Specify the number of UltraServers to search for.

Reserve the best training plan

After reviewing the available training plan offerings that best match your requirements, you can reserve a specific plan by calling the [`CreateTrainingPlan`](#) API operation. When created, the plan initially enters a `Pending` state and remains there until the reservation process is complete.

The response to the API call returns a training plan Amazon Resource Name (ARN). Make a note of this ARN for tracking and monitoring purposes later on. The training plan reservation is fulfilled asynchronously in the backend. The payment for the total amount is automatically collected as part of the fulfillment process. Once the payment transaction is completed and the requested reserved capacities are secured, the training plan is set to the `Scheduled` state, and is ready for scheduling.

> **Important**
>
> - Training plans cannot be modified once purchased.
> - Training plans cannot be shared across AWS accounts or within your AWS Organization.

The following example uses the an AWS CLI command to request a specific training plan, passing the plan ID as a parameter.

```
aws sagemaker create-training-plan \
--training-plan-offering-id "tpo-SHA-256-hash-value" \
--training-plan-name "name" \
```

This JSON document is a sample response from the SageMaker training plans API. The response contains the Amazon Resource Name (ARN) of the training plan that has been successfully created.

> **Note**
>
> The training plan remains in a `Pending` status until the fulfillment process is complete.

```
{
   "TrainingPlanArn":"arn:aws:sagemaker:us-east-1:123456789123:training-plan/large-models-fine-tuning"
}
```

The following sections define the mandatory and optional input request parameters for the `CreateTrainingPlan` API operation.

Required parameters

When calling CreateTrainingPlan API to reserve a particular training plan, you must provide the following values:

- TrainingPlanOfferingId: The ID of the plan you are choosing. You can retrieve the ID of a plan offering in the response of your SearchTrainingPlanOfferings API call. Its format should start with pto-*.
- TrainingPlanName: The name of the plan you are creating.

List training plans

You can list all the training plans that have been created in your AWS account and Region by calling the ListTrainingPlans API.

The following example uses an AWS CLI command to retrieve the list of your training plans.

```
aws sagemaker list-training-plans \
--start-time-after "2024-09-26T00:00:01.000Z"
```

This JSON document is a sample response from the SageMaker training plans API. The response provides details about one training plan that has been successfully created and reserved.

```
{
   "TrainingPlanSummaries": [
      {
         "AvailableInstanceCount": 2,
         "CurrencyCode": "USD",
         "DurationHours": 48,
         "DurationMinutes": 0,
         "EndTime": "2024-09-28T04:30:00-07:00",
         "InUseInstanceCount": 2,
         "ReservedCapacitySummaries": [
            {
               "AvailabilityZone": "string",
               "DurationHours": 48,
               "DurationMinutes": 0,
               "EndTime": "2024-09-28T04:30:00-07:00",
               "InstanceType": "ml.p5.48xlarge",
               "ReservedCapacityArn": "arn:aws:sagemaker:us-
east-1:123456789123:reserved-capacity/large-models-fine-tuning-rc1",
               "StartTime": "2024-09-26T04:30:00-07:00",
               "Status": "Scheduled",
               "TotalInstanceCount": 4,
               "UltraServerCount": 4,
               "UltraServerType": "ml.p6e-gb200.36xlarge"
            }
         ],
         "StartTime": "2024-09-26T04:30:00-07:00",
         "Status": "Scheduled",
         "StatusMessage": "Payment confirmed, training plan scheduled."
         "TargetResources": [ "training-job" ],
         "TotalInstanceCount": 4,
         "TotalUltraServerCount": 4,
         "TrainingPlanArn": "arn:aws:sagemaker:us-east-1:123456789123:training-plan/
large-models-fine-tuning",
         "TrainingPlanName": "large-models-fine-tuning",
         "UpfrontFee": "xxxx.xx"
      }
   ]
}
```

The following sections provide details of some optional parameters that you can pass to your `ListTrainingPlans` API request.

Optional parameters

The following sections provide details on some optional parameters that you can pass to your `ListTrainingPlans` API request.

- `StartTimeAfter`: The start time of the actual time range of the listed plans, specified as a `timestamp` or an `ISO 8601 date/time`.
- `StartTimeBefore`: The end time of the actual time range of the listed plans, specified as a `timestamp` or an `ISO 8601 date/time`.
- `Filters`: Criteria used to filter the results, with up to 5 Name-Value pairs where "Name" is the name of a field of a TrainingPlanSummary and "Value" is the value to consider for the filter. For example `Name=Status,Value=Active`.

The following example uses an AWS CLI command to retrieve your list of training plans, using some of the optional parameters described above.

```
aws sagemaker list-training-plans --max-results 10 --sort-by StartTime --sort-order
 Descending --start-time-after 13000000 --filters Name=Status,Value=Active
```

View training plan details

To monitor the status or retrieve details of a training plan, you can use the `DescribeTrainingPlan` API. The API response includes a `Status` field, which reflects the current state of the training plan:

- If the plan purchase fails, the status is set to `Failed`.
- Upon successful payment, the status transitions from `Pending` to `Scheduled`, based on the plan's start date.
- When the plan reaches its start date, the status changes to `Active`.
- For plans with multiple discontinuous reserved capacities, the status reverts to `Scheduled` between active periods, until the start date of the next reserved capacity.
- After the plan's end date, the status becomes `Expired`.

Once the status is `Scheduled`, you can utilize the capacity reserved in the plan for your SageMaker training jobs or HyperPod cluster workloads.

> **Note**
>
> - Training jobs associated with the plan remain in `Pending` status until the plan becomes `Active`.
> - For HyperPod clusters using a training plan for compute capacity, the instance group status appears as `InService` once created.

The following example uses an AWS CLI command to retrieve the details of a training plan by its name.

```
aws sagemaker describe-training-plan \
--training-plan-name "name"
```

This JSON document is a sample response from the SageMaker training plans API. This response provides details about a training plan that has been successfully created.

```
      {
         "AvailableInstanceCount": 2,
         "CurrencyCode": "USD",
         "DurationHours": 48,
         "DurationMinutes": 0,
         "EndTime": "2024-09-28T04:30:00-07:00",
         "InUseInstanceCount": 2,
         "ReservedCapacitySummaries": [
            {
               "AvailabilityZone": "string",
               "DurationHours": 48,
               "DurationMinutes": 0,
               "EndTime": "2024-09-28T04:30:00-07:00",
               "InstanceType": "ml.p5.48xlarge",
               "ReservedCapacityArn": "arn:aws:sagemaker:us-
east-1:123456789123:reserved-capacity/large-models-fine-tuning-rc1",
               "StartTime": "2024-09-26T04:30:00-07:00",
               "Status": "Scheduled",
               "TotalInstanceCount": 4,
               "UltraServerCount": 4,
               "UltraServerType": "ml.p6e-gb200.36xlarge"
            }
         ],
         "StartTime": "2024-09-26T04:30:00-07:00",
         "Status": "Scheduled",
         "StatusMessage": "Payment confirmed, training plan scheduled."
         "TargetResources": [ "training-job" ],
         "TotalInstanceCount": 4,
         "TotalUltraServerCount": 4,
         "TrainingPlanArn": "arn:aws:sagemaker:us-east-1:123456789123:training-plan/
large-models-fine-tuning",
         "TrainingPlanName": "large-models-fine-tuning",
         "UpfrontFee": "xxxx.xx"
      }
```

The following sections define the mandatory input request parameter for the `DescribeTrainingPlan` API operation.

Required parameters

- `TrainingPlanName`: The name of the training plan you want to describe.

Training plans utilization for SageMaker training jobs

You can use a SageMaker training plans for your training jobs by specifying the plan of your choice when creating a training job.

Note

The training plan must be in the `Scheduled` or `Active` status to be used by a training job.

If the required capacity is not immediately available for a training job, the job waits until it becomes available, or until the `StoppingCondition` is met, or the job has been `Pending` for capacity for 2 days, whichever comes first. If the stopping condition is met, the job is stopped. If a job has been pending for 2 days, it is terminated with an `InsufficientCapacityError`.

Important

Reserved Capacity termination process: You have full access to all reserved instances until 30 minutes before the Reserved Capacity end time. When there are 30 minutes remaining in your Reserved Capacity, SageMaker training plans begin the process of terminating any running instances within that Reserved Capacity.
To ensure you don't lose progress due to these terminations, we recommend checkpointing your training jobs.

Checkpoint your training job

When using SageMaker training plans for your SageMaker training jobs, ensure to implement checkpointing in your training script. This allows you to save your training progress before a Reserved Capacity expires. Checkpointing is especially important when working with reserved capacities, as it enables you to resume training from the last saved point if your job is interrupted between two reserved capacities or when your training plan reaches its end date.

To achieve this, you can use the `SAGEMAKER_CURRENT_CAPACITY_BLOCK_EXPIRATION_TIMESTAMP` environment variable. This variable helps determine when to initiate the checkpointing process. By incorporating this logic into your training script, you ensure that your model's progress is saved at appropriate intervals.

Here's an example of how you can implement this checkpointing logic in your Python training script:

```
import os
import time
from datetime import datetime, timedelta

def is_close_to_expiration(threshold_minutes=30):
    # Retrieve the expiration timestamp from the environment variable
    expiration_time_str =
 os.environ.get('SAGEMAKER_CURRENT_CAPACITY_BLOCK_EXPIRATION_TIMESTAMP', '0')

    # If the timestamp is not set (default '0'), return False
    if expiration_time_str == '0':
        return False

    # Convert the timestamp string to a datetime object
    expiration_time = datetime.fromtimestamp(int(expiration_time_str))

    # Calculate the time difference between now and the expiration time
    time_difference = expiration_time - datetime.now()

    # Return True if we're within the threshold time of expiration
    return time_difference < timedelta(minutes=threshold_minutes)

def start_checkpointing():
    # Placeholder function for checkpointing logic
    print("Starting checkpointing process...")
    # TODO: Implement actual checkpointing logic here
    # For example:
    # - Save model state
    # - Save optimizer state
    # - Save current epoch and iteration numbers
    # - Save any other relevant training state

# Main training loop
num_epochs = 100
final_checkpointing_done = False
for epoch in range(num_epochs):
    # TODO: Replace this with your actual training code
    # For example:
    # - Load a batch of data
    # - Forward pass
    # - Calculate loss
    # - Backward pass
    # - Update model parameters

    # Check if we're close to capacity expiration and haven't done final checkpointing
    if not final_checkpointing_done and is_close_to_expiration():
        start_checkpointing()
        final_checkpointing_done = True

    # Simulate some training time (remove this in actual implementation)
    time.sleep(1)
print("Training completed.")
```

Note

- Training job provisioning follows a First-In-First-Out (FIFO) order, but a smaller cluster job created later might be assigned capacity before a larger cluster job created earlier, if the larger job cannot be fulfilled.
- SageMaker training managed warm-pool is compatible with SageMaker training plans. For cluster re-use, you must provide identical `TrainingPlanArn` values in subsequent `CreateTrainingJob` requests to reuse the same cluster.

Topics

- Create a training job using the SageMaker AI console
- Create a training job using the API, AWS CLI, SageMaker SDK

Create a training job using the SageMaker AI console

You can use a SageMaker training plans for your training jobs using the SageMaker AI UI. When creating a training job, the available plans are suggested to you if your instance choice and region matches the available plans.

To create a training job using a training plan's reserved capacity in the SageMaker console:

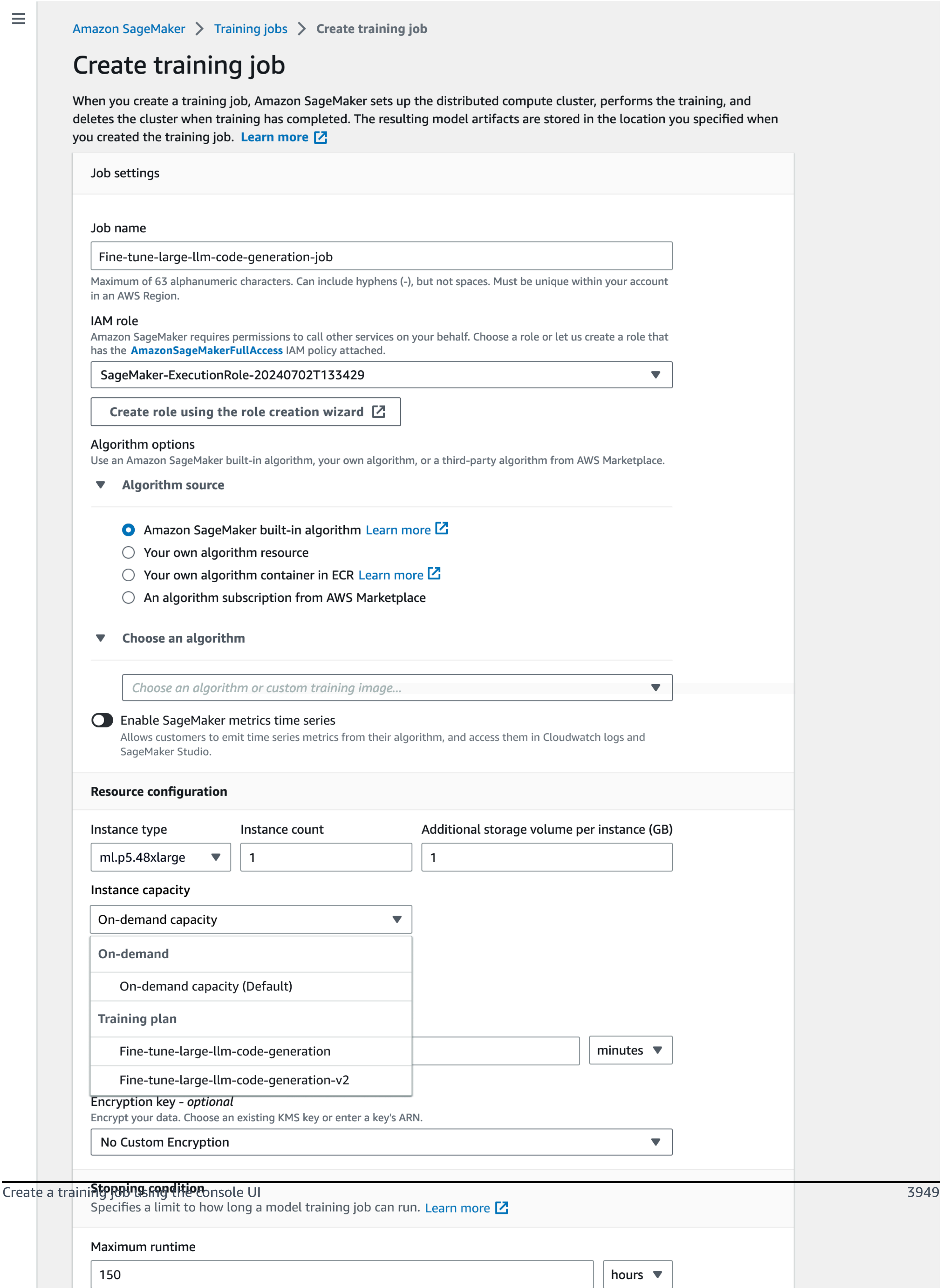
1. Navigate to the SageMaker AI console at https://console.aws.amazon.com/sagemaker/.
2. In the left navigation pane, choose **Training**, and then **Training jobs**.

3. Choose the **Create training job** button.

4. When configuring the resources for your training job, look for the **Instance capacity** section. If there are plans available that match your chosen instance type and region, they are displayed here. Select a training plan that aligns with your compute capacity needs.

   If no suitable plans are available, you can either adjust your instance type or region, or proceed without using a training plan.

5. After selecting a training plan (or choosing to proceed without one), complete the rest of your training job configuration and choose **Create training job** to start the process.



Create training job

When you create a training job, Amazon SageMaker sets up the distributed compute cluster, performs the training, and deletes the cluster when training has completed. The resulting model artifacts are stored in the location you specified when you created the training job. Learn more

Job settings

Job name

Fine-tune-large-llm-code-generation-job

Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

IAM role

Amazon SageMaker requires permissions to call other services on your behalf. Choose a role or let us create a role that has the **AmazonSageMakerFullAccess** IAM policy attached.

SageMaker-ExecutionRole-20240702T133429

Algorithm options

Use an Amazon SageMaker built-in algorithm, your own algorithm, or a third-party algorithm from AWS Marketplace.

Algorithm source

● Amazon SageMaker built-in algorithm Learn more

○ Your own algorithm resource

○ Your own algorithm container in ECR Learn more

○ An algorithm subscription from AWS Marketplace

Choose an algorithm

Choose an algorithm or custom training image...

Enable SageMaker metrics time series

Allows customers to emit time series metrics from their algorithm, and access them in Cloudwatch logs and SageMaker Studio.

Resource configuration

Instance type	Instance count	Additional storage volume per instance (GB)
ml.p5.48xlarge	1	1

Instance capacity

On-demand capacity

On-demand

On-demand capacity (Default)

Training plan

Fine-tune-large-llm-code-generation

Fine-tune-large-llm-code-generation-v2

minutes

Encryption key - *optional*

Encrypt your data. Choose an existing KMS key or enter a key's ARN.

No Custom Encryption

Stopping condition

Specifies a limit to how long a model training job can run. Learn more

Maximum runtime

150 hours

Review and launch your job. Your job starts running as soon as the training plan becomes `Active`, pending capacity.

Create a training job using the API, AWS CLI, SageMaker SDK

To use SageMaker training plans for your SageMaker training job, specify the `TrainingPlanArn` parameter of the desired plan in the `ResourceConfig` when calling the [`CreateTrainingJob`](CreateTrainingJob) API operation. You can use exactly one plan per job.

> ⚠ **Important**
>
> The `InstanceType` field set in the `ResourceConfig` section of the `CreateTrainingJob` request must match the`InstanceType` of your training plan.

Run a training job on a plan using the CLI

The following example demonstrates how to create a SageMaker training job and associate it with a provided training plan using the `TrainingPlanArn` attribute in the `create-training-job` AWS CLI command.

For more information about how to create a training job using the AWS CLI CreateTrainingJob command, see [`create-training-job`](create-training-job).

```
# Create a training job
aws sagemaker create-training-job \
  --training-job-name training-job-name \
  ...

  --resource-config '{
        "InstanceType": "ml.p5.48xlarge",
        "InstanceCount": 8,
        "VolumeSizeInGB": 10,
        "TrainingPlanArn": "training-plan-arn"
        }
    }' \
    ...
```

This AWS CLI example command creates a new training job in SageMaker AI passing a training plan in the `--resource-config` argument.

```
aws sagemaker create-training-job \
  --training-job-name job-name \
  --role-arn arn:aws:iam::111122223333:role/DataAndAPIAccessRole \
  --algorithm-specification '{"TrainingInputMode": "File","TrainingImage":
 "111122223333.dkr.ecr.us-east-1.amazonaws.com/algo-image:tag", "ContainerArguments":
 [" "]}' \
  --input-data-config '[{"ChannelName":"training","DataSource":
{"S3DataSource":{"S3DataType":"S3Prefix","S3Uri":"s3://bucketname/
input","S3DataDistributionType":"ShardedByS3Key"}}}]' \
  --output-data-config '{"S3OutputPath": "s3://bucketname/output"}' \
  --resource-config
 '{"VolumeSizeInGB":10,"InstanceCount":4,"InstanceType":"ml.p5.48xlarge",
 "TrainingPlanArn" : "arn:aws:sagemaker:us-east-1:111122223333:training-plan/plan-
name"}' \
  --stopping-condition '{"MaxRuntimeInSeconds": 1800}' \
  --region us-east-1
```

After creating the training job, you can verify that it was properly assigned to the training plan by calling the `DescribeTrainingJob` API.

```
aws sagemaker describe-training-job --training-job-name training-job-name
```

Run a training job on a plan using the SageMaker AI Python SDK

Alternatively, you can create a training job associated with a training plan using the SageMaker Python SDK.

If you are using the SageMaker Python SDK from JupyterLab in Studio to create a training job, ensure that the execution role used by the space running your JupyterLab application has the required permissions to use SageMaker training plans. To learn about the required permissions to use SageMaker training plans, see the section called "IAM for SageMaker training plans".

The following example demonstrates how to create a SageMaker training job and associate it with a provided training plan using the `training_plan` attribute in the `Estimator` object when using the SageMaker Python SDK.

For more information on the SageMaker Estimator, see Use a SageMaker estimator to run a training job.

```
import sagemaker
import boto3
from sagemaker import get_execution_role
from sagemaker.estimator import Estimator
from sagemaker.inputs import TrainingInput

# Set up the session and SageMaker client
session = boto3.Session()
region = session.region_name
sagemaker_session = session.client('sagemaker')

# Get the execution role for the training job
role = get_execution_role()

# Define the input data configuration
trainingInput = TrainingInput(
    s3_data='s3://input-path',
    distribution='ShardedByS3Key',
    s3_data_type='S3Prefix'
)

estimator = Estimator(
    entry_point='train.py',
    image_uri="123456789123.dkr.ecr.{}.amazonaws.com/image:tag",
    role=role,
    instance_count=4,
    instance_type='ml.p5.48xlarge',
    training_plan="training-plan-arn",
    volume_size=20,
    max_run=3600,
    sagemaker_session=sagemaker_session,
    output_path="s3://output-path"
)

# Create the training job
estimator.fit(inputs=trainingInput, job_name=job_name)
```

After creating the training job, you can verify that it was properly assigned to the training plan by calling the `DescribeTrainingJob` API.

```
# Check job details
sagemaker_session.describe_training_job(TrainingJobName=job_name)
```

Training plans utilization for Amazon SageMaker HyperPod clusters

To use SageMaker training plans for your Amazon SageMaker HyperPod cluster, you specify the training plan you want to use at the cluster instance level when creating or updating your cluster.

Note

- The training plan must be in the `Scheduled` or `Active` status to be used by an HyperPod cluster.
- Ensure the cluster configuration aligns with the Availability Zone (AZ) specified in your training plan.

  For VPC setup, resource location, and security group configuration, refer to the section called "Setting up SageMaker HyperPod with a custom Amazon VPC" in the SageMaker HyperPod documentation.

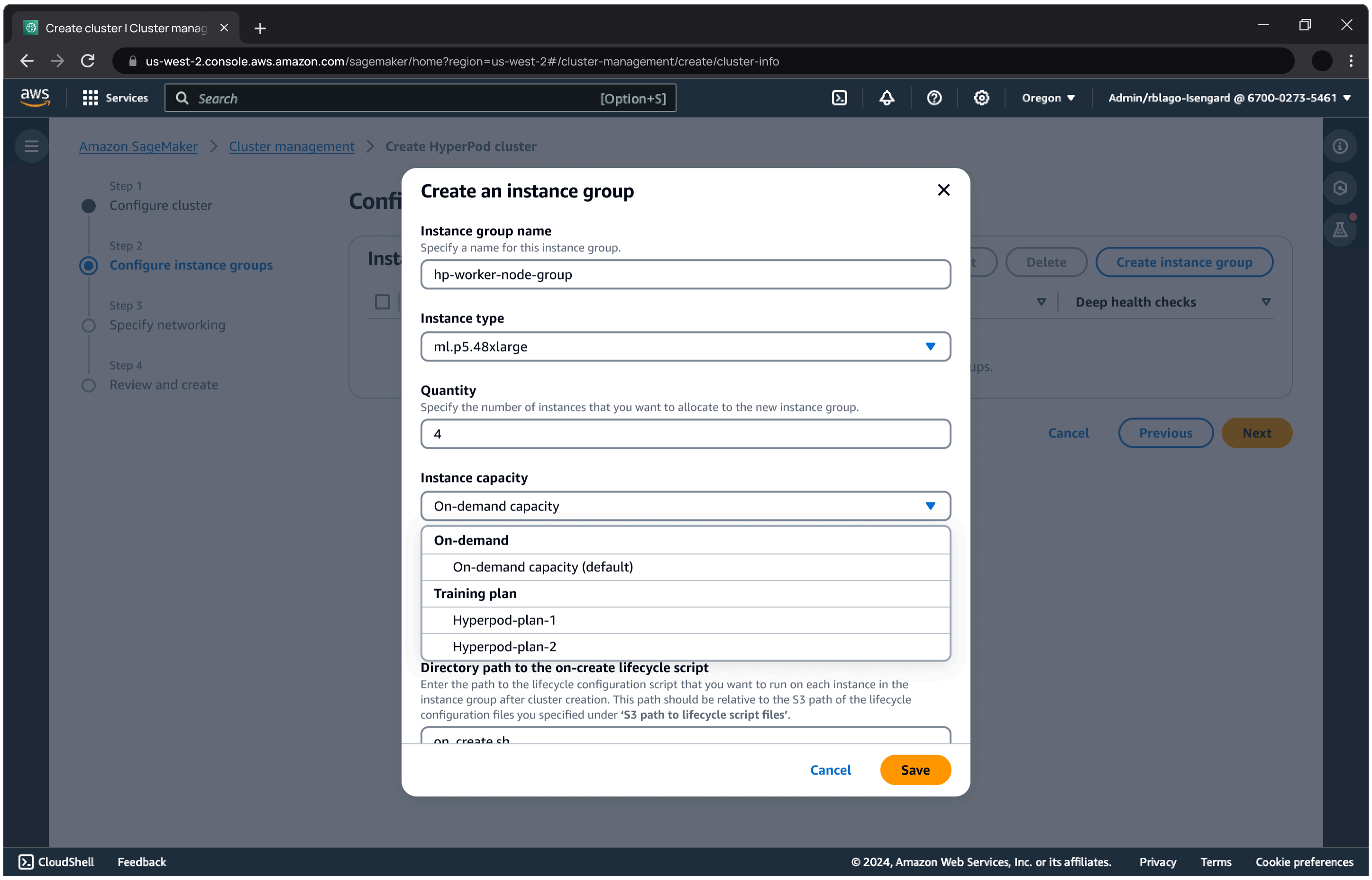
  If setting up HyperPod with Amazon FSx for Lustre, learn about Region and AZ selection, review VPC configuration requirements, and understand AZ alignment best practices in the section called "(Optional) Setting up SageMaker HyperPod with Amazon FSx for Lustre".
- You can select a plan for each of your instance groups. However, we do not recommend using a training plan for the primary instance group of a cluster, as primary nodes require continuous, stable resources that don't align with the fixed duration and potentially discontinuous nature of training plan capacities.

Topics

Create a SageMaker HyperPod cluster on training plans using the SageMaker AI console

To create an SageMaker HyperPod cluster using training plans from the SageMaker AI console UI, follow these steps:

1. Navigate to the SageMaker AI console at https://console.aws.amazon.com/sagemaker/.
2. In the left navigation pane, choose **Hyperpod**, and then **Create cluster**.
3. When configuring an instance group, you can select a plan that aligns with your compute capacity needs.



Create cluster | Cluster manag
us-west-2.console.aws.amazon.com/sagemaker/home?region=us-west-2#/cluster-management/create/cluster-info
Services
Search
[Option+S]
Oregon
Admin/rblago-Isengard @ 6700-0273-5461
Amazon SageMaker > Cluster management > Create HyperPod cluster
Step 1
Configure cluster
Step 2
Configure instance groups
Step 3
Specify networking
Step 4
Review and create
Delete
Create instance group
Deep health checks
Cancel
Previous
Next
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Create an instance group

Instance group name
Specify a name for this instance group.
hp-worker-node-group

Instance type
ml.p5.48xlarge

Quantity
Specify the number of instances that you want to allocate to the new instance group.
4

Instance capacity
On-demand capacity

On-demand
On-demand capacity (default)
Training plan
Hyperpod-plan-1
Hyperpod-plan-2

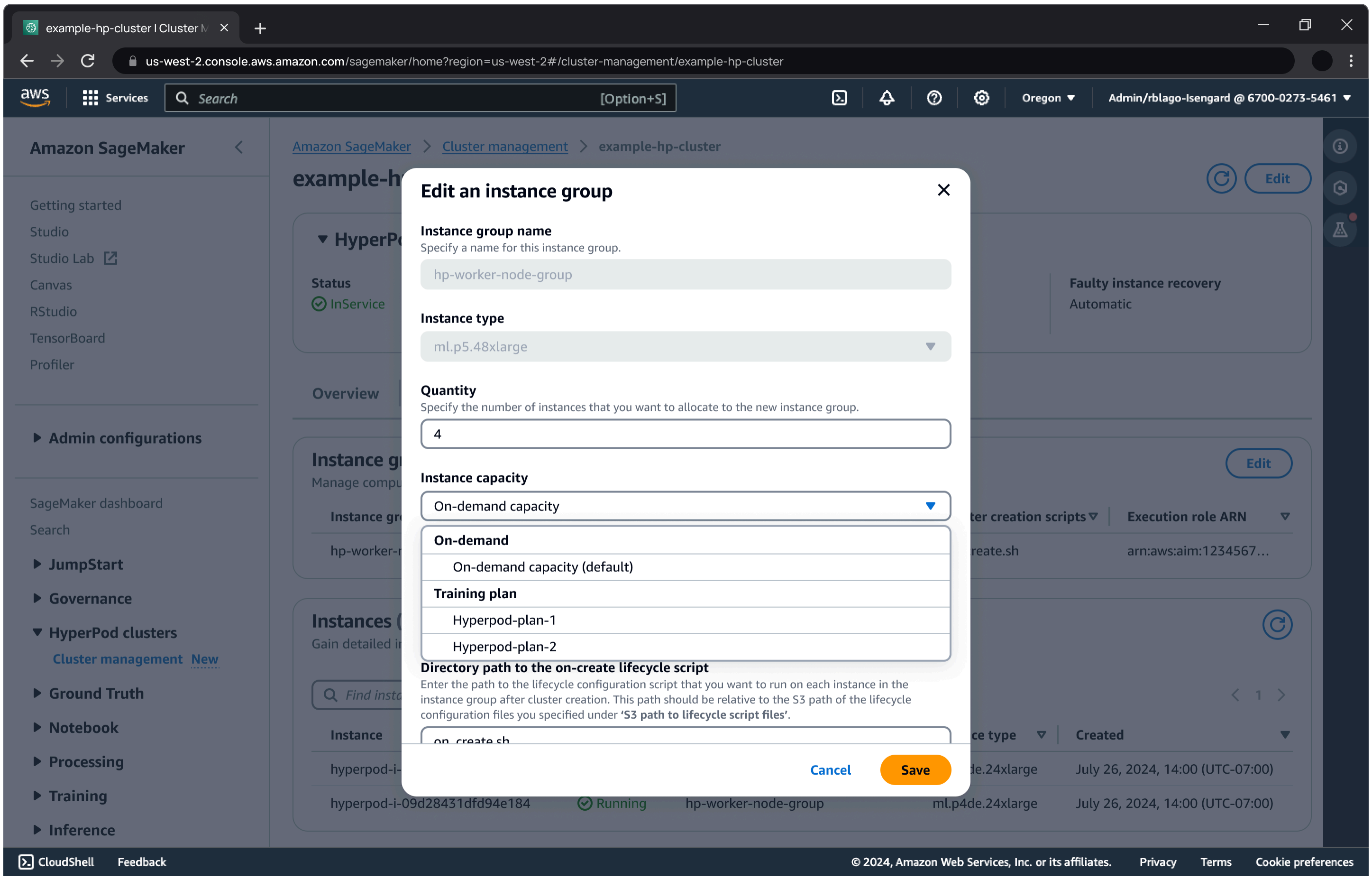
Directory path to the on-create lifecycle script
Enter the path to the lifecycle configuration script that you want to run on each instance in the instance group after cluster creation. This path should be relative to the S3 path of the lifecycle configuration files you specified under **'S3 path to lifecycle script files'**.
on_create.sh

Review and create your cluster. Instance groups using a training plan scale up to the specified target instance count when the training plan becomes `Active`, subject to available capacity. Thirty minutes before each Reserved Capacity period ends, the instance group begins scaling down to zero instances. This scaled-down state persists until the next Reserved Capacity period begins or the plan ends. Throughout this process, an healthy instance group maintains an `InService` status after its initial creation, regardless of the current instance count.

Update a SageMaker HyperPod cluster on training plans using the SageMaker AI console

You can update, remove, or add a training plan to an existing SageMaker HyperPod cluster using the SageMaker AI console UI. To update the instance group of an SageMaker HyperPod cluster, follow these steps:

1. Navigate to the SageMaker AI console at https://console.aws.amazon.com/sagemaker/.
2. In the left navigation pane, choose **Hyperpod**.
3. Navigate to the cluster's details page by following the hyperlink associated with the cluster name.
4. When configuring an instance group, you can update your plan to align with your new compute capacity needs.



Edit an instance group

Instance group name
Specify a name for this instance group.
hp-worker-node-group

Instance type
ml.p5.48xlarge

Quantity
Specify the number of instances that you want to allocate to the new instance group.
4

Instance capacity
On-demand capacity

- On-demand
  - On-demand capacity (default)
- Training plan
  - Hyperpod-plan-1
  - Hyperpod-plan-2

Directory path to the on-create lifecycle script
Enter the path to the lifecycle configuration script that you want to run on each instance in the instance group after cluster creation. This path should be relative to the S3 path of the lifecycle configuration files you specified under '**S3 path to lifecycle script files**'.
on_create.sh

Instance			ce type	Created
hyperpod-i-			de.24xlarge	July 26, 2024, 14:00 (UTC-07:00)
hyperpod-i-09d28431dfd94e184	Running	hp-worker-node-group	ml.p4de.24xlarge	July 26, 2024, 14:00 (UTC-07:00)

Review and update your cluster.

Create a SageMaker HyperPod cluster on training plans using the SageMaker API, or AWS CLI

To use SageMaker training plans for your Amazon SageMaker HyperPod cluster, specify the ARN of the training plan you want to use in the `TrainingPlanArn` parameter of the `ClusterInstanceGroupSpecification` when calling the `CreateCluster` API operation.

Ensure that the subnet associated with the designated AZ of your plan is included in the `VPCConfig` of your cluster configuration. You can retrieve the `AvailabilityZone` of a training plan in the response of a DescribeTrainingPlan API call.

The following sample illustrates how to create a new SageMaker HyperPod cluster and provide an instance group with a training plan in the `--instance-groups` attribute of the `create-cluster` AWS CLI command.

```
# Create a cluster
```

```
aws sagemaker create-cluster \
  --cluster-name cluster-name \
  --instance-groups '[ \
        { \
            "InstanceCount": 1,\
            "InstanceGroupName": "controller-nodes",\
            "InstanceType": "ml.t3.xlarge",\
            "LifeCycleConfig": {"SourceS3Uri": source_s3_uri, "OnCreate":
 "on_create.sh"},\
            "ExecutionRole": "arn:aws:iam::customer_account_id:role/execution_role",\
            "ThreadsPerCore": 1,\
        },\
        { \
            "InstanceCount": 2, \
            "InstanceGroupName": "worker-nodes",\
            "InstanceType": "p4d.24xlarge",\
            "LifeCycleConfig": {"SourceS3Uri": source_s3_uri, "OnCreate":
 "on_create.sh"},\
            "ExecutionRole": "arn:aws:iam::customer_account_id}:role/execution_role}",\
            "ThreadsPerCore": 1,\
            "TrainingPlanArn": training_plan_arn,\
        }]'
```

For information about how to create an HyperPod cluster using the AWS CLI, see `create-cluster`.

After creating the cluster, you can verify that your instance group was properly assigned capacity from the training plan by calling the `DescribeCluster` API.

```
aws sagemaker describe-cluster --cluster-name cluster-name
```

Update a SageMaker HyperPod cluster on training plans using the SageMaker API, or AWS CLI

You can add, update, or remove a training plan by updating the instance group of an existing cluster using the `update-cluster` AWS CLI command. The following sample illustrates how to update a SageMaker HyperPod cluster and provide an instance group with a new training plan.

```
# Update a cluster
aws sagemaker update-cluster \
  --cluster-name cluster-name \
  --instance-groups '[ \
        { \
            "InstanceCount": 1,\
            "InstanceGroupName": "controller-nodes",\
            "InstanceType": "ml.t3.xlarge",\
            "LifeCycleConfig": {"SourceS3Uri": source_s3_uri, "OnCreate":
 "on_create.sh"},\
            "ExecutionRole": "arn:aws:iam::customer_account_id:role/execution_role",\
            "ThreadsPerCore": 1,\
        },\
        { \
            "InstanceCount": 2, \
            "InstanceGroupName": "worker-nodes",\
            "InstanceType": "p4d.24xlarge",\
            "LifeCycleConfig": {"SourceS3Uri": source_s3_uri, "OnCreate":
 "on_create.sh"},\
            "ExecutionRole": "arn:aws:iam::customer_account_id}:role/execution_role}",\
            "ThreadsPerCore": 1,\
            "TrainingPlanArn": training_plan_arn,\
        },\
        {\
            "InstanceCount": 1,\
            "InstanceGroupName": "worker-nodes-2",\
            "InstanceType": "p4d.24xlarge",\
            "LifeCycleConfig": {"SourceS3Uri": source_s3_uri, "OnCreate":
 "on_create.sh"},\
            "ExecutionRole": "arn:aws:iam::customer_account_id:role/execution_role",\
            "ThreadsPerCore": 1,\
            "TrainingPlanArn": training_plan_arn,\
        }\
    ]'
```

View SageMaker training plans quotas using the AWS management console

⚠ **Important**

- For pricing information about SageMaker training plans, see the Amazon SageMaker Pricing page. Navigate to the **Amazon SageMaker HyperPod flexible training plans** section under **On-Demand Pricing**. Choose your desired Region to view available instance types and their corresponding prices.

- Make sure that your Training Jobs or HyperPod service quotas allow a maximum number of instances per instance type that exceeds the number of instances specified in your plan.

You can view the current quotas and limits for SageMaker training plans using the AWS Management Console.

To search for a specific quota value:

1. Open the Service Quotas console.
2. In the left navigation pane, choose **AWS services**.
3. From the AWS services list, search for and select **Amazon SageMaker AI**.
4. In the **Service quotas** list, you can see the service quota name, applied value (if it's available), AWS default quota, and whether the quota value is adjustable.

To find specific quotas, you can use the search bar at the top of the **Service quotas** list. Type the `Limit Name` of the quota you are searching for. For example, to find the quota for the number of training plans per region, you would type **`training-plan-total_count`** in the search bar.

The following table outlines the quota limit names for SageMaker training plans.

SageMaker training plans quota limits

Limit Name	Display Name
training-plan-total_count	Number of training plans per Region
reserved-capacity-ml-p4d-24xlarge	Number of ml.p4d.24xlarge instances in reserved capacity across training plans per Region
reserved-capacity-ml-p5-48xlarge	Number of ml.p5.48xlarge instances in reserved capacity across training plans per Region

Limit Name	Display Name
reserved-capacity-ml-p5e-48xlarge	Number of ml.p5e.48xlarge instances in reserved capacity across training plans per Region
reserved-capacity-ml-p5en-48xlarge	Number of ml.p5en.48xlarge instances in reserved capacity across training plans per Region
reserved-capacity-ml-trn1-32xlarge	Number of ml-trn1-32xlarge instances in reserved capacity across training plans per Region
reserved-capacity-ml-trn2-48xlarge	Number of ml.trn2.48xlarge instances in reserved capacity across training plans per Region

If you need higher limits for your SageMaker training plans, you may be able to request a quota increase. The ability to increase a quota depends on whether it's adjustable, which you can see in the **Service quotas** console.

To request a quota increase:

1. Navigate to the specific quota in the **Service quotas** console.
2. If the quota is adjustable, you can request a quota increase at either the account level or resource level based on the value listed in the **Adjustability** column.
3. For **Increase quota value**, enter the new value. The new value must be greater than the current value.
4. Choose **Request**.
5. Quota increase requests are subject to review and approval by AWS. To view any pending or recently resolved requests in the console, navigate to the **Request history** tab from the service's details page, or choose **Dashboard** from the navigation pane. For pending requests, choose the status of the request to open the request receipt. The initial status of a request is `Pending`. After the status changes to `Quota requested`, you see the case number with AWS Support. Choose the case number to open the ticket for your request.

To learn more about requesting a quota increase in general, see Requesting a quota increase in the *AWS Service Quotas User Guide*.

Release notes

See the following release notes to track the latest updates for SageMaker training plans.

Amazon SageMaker training plans Release Notes: December 04, 2024

New Features

- Launched Amazon SageMaker training plans at AWS re:Invent 2024.