Step-by-Step Guide: Creating an AWS SageMaker Ground Truth Labeled Workforce and Datasets

Introduction

Amazon SageMaker Ground Truth is a powerful service within AWS SageMaker that helps you build high-quality training datasets for machine learning models by facilitating data labeling. It supports various labeling tasks like image classification, text classification, bounding boxes, and more. You can use different types of workforces: private (your own team), vendor-managed, or public (via Amazon Mechanical Turk).

In this blog, we'll focus on creating a **private workforce** (recommended for sensitive data or testing) and setting up a **labeling job** to label a dataset. We'll use a simple text classification example for illustration, but the process is similar for other data types like images.

Prerequisites:

- An AWS account with access to SageMaker.
- Appropriate IAM permissions (e.g., AmazonSageMakerFullAccess policy, plus Cognito permissions for workforce creation).
- Data uploaded to an Amazon S3 bucket (e.g., text files or images).
- For this example, create two S3 buckets: one for input data (e.g., your-input-bucket) and one for output (e.g., your-output-bucket).

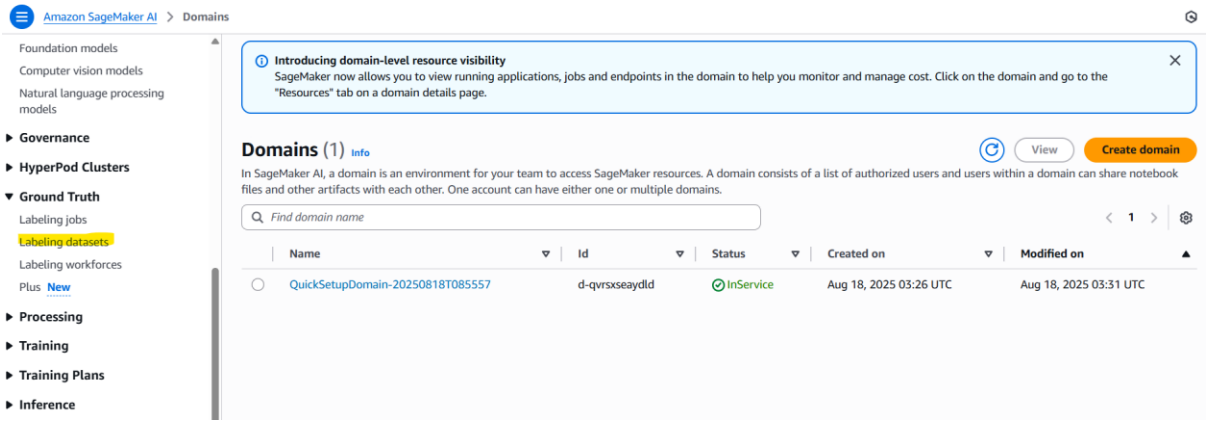
Part 1: Creating a Private Workforce

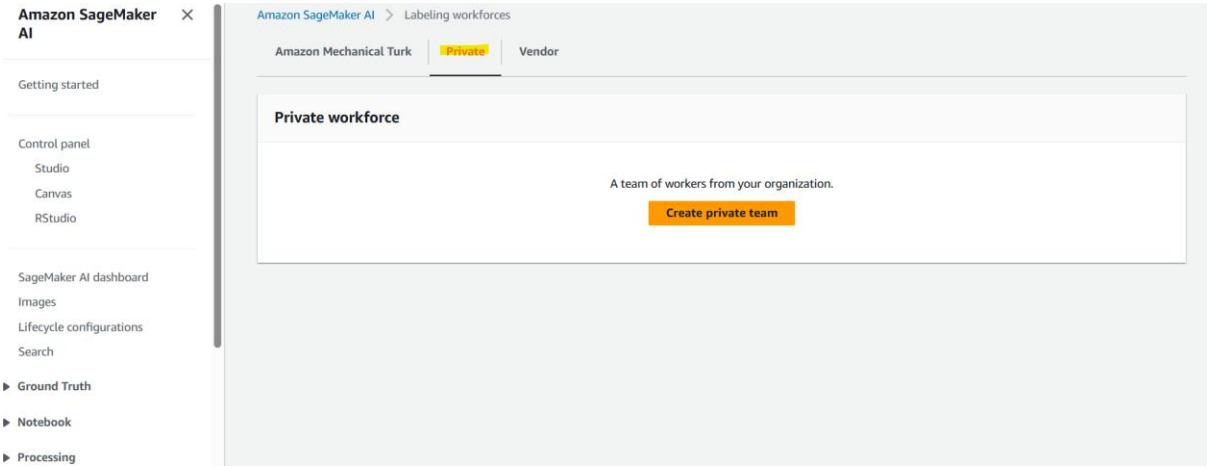
A private workforce consists of your own team members (e.g., employees or contractors) who label data via a secure portal. It's authenticated using Amazon Cognito.

There are two main ways to create it: during a labeling job setup or via the Labeling workforces page. We'll cover the latter for independence.

Step 1: Navigate to the SageMaker Console

- Open the AWS Management Console and go to SageMaker: https://console.aws.amazon.com/sagemaker/.
- In the left navigation pane, under **Ground Truth**, select **Labeling workforces**.





Step 2: Select the Private Tab and Create a Team

- Switch to the **Private** tab.
- Click **Create private team**.
- Choose **Invite new workers by email**.

Step 3: Invite Workers

- In the email addresses box, paste or type up to 50 email addresses (comma-separated). These are case-sensitive.
- Enter your **Organization name** (used in invitation emails).
- Enter a **Contact email** for workers to report issues.
- Optionally, select an SNS topic to notify workers about new jobs (email notifications for Ground Truth jobs only).

Step 4: Create the Team

- Click **Create private team**.
- Refresh the page to see the **Private workforce summary**, which includes:

    - Cognito user pool details.
    - List of work teams.
    - List of workforce members.
- Invited workers will receive an email with a login link and temporary password. They must change the password on first login.

Amazon SageMaker AI > Labeling workforces > Create private team

Create private team

Private team creation

- (●) Create a private team with AWS Cognito
  Create a private work team by sending email invitations to new workers or importing workers from existing Amazon Cognito user groups.
- ( ) Create a private team with OpenID Connect (OIDC)
  Create a private work team with your own identity provider (IdP). Your IdP must support OIDC user groups.

Team details

Team name
Give your work team a descriptive name. This name can't be changed later.

Image-Classification-Testing-Team

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

Add workers Info

Add workers to your private work team by adding worker email addresses or importing workers from existing Amazon Cognito user groups.

- (●) Invite new workers by email
- ( ) Import workers from existing Amazon Cognito user groups

Email addresses
We send an invitation with instructions to each of the worker email addresses that you add here.

user1@gmail.com, user2@gmail.com, user3@gmail.com

Use a comma between addresses. You can add up to 50 workers.

Organization name
We use this information to customize the email that we send to the workers.

TECHBLOGGER-ORGANIZATION

Contact email
Workers use this address to report issues related to the task.

support-contact@gmail.com

ⓘ We send an email with the login details to all the workers added to your team.

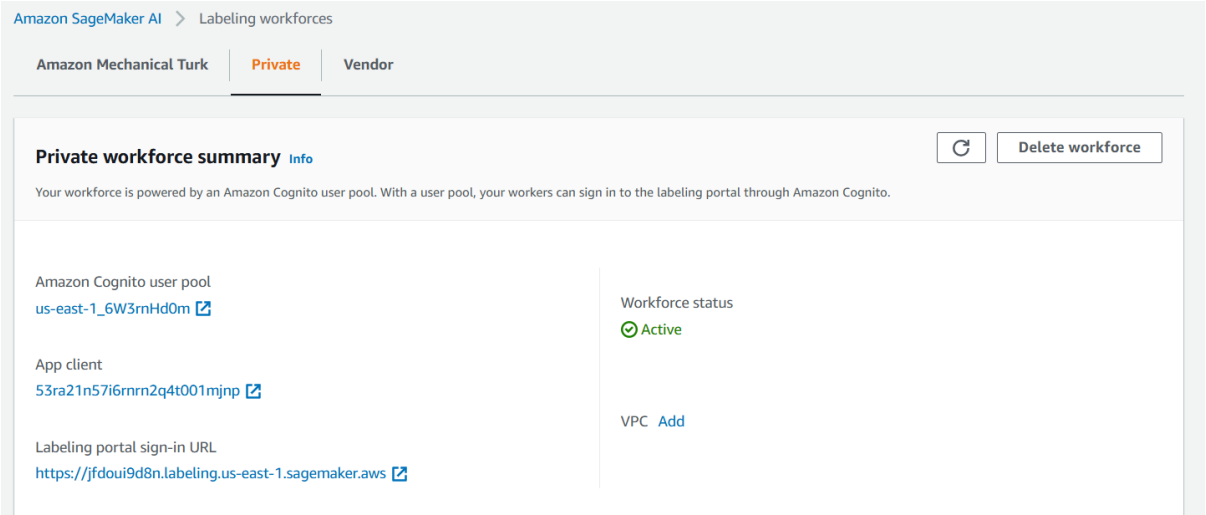
Email Invitation
We send an email with the login details to all the workers added to your team.

Preview invitation

Enable Notifications

SNS topic - *optional*
Configuring SNS topic enables your work team to receive notifications on available work. Learn more

Once created, you will be able to view something like the following,



Step 5: Manage the Workforce (Optional)

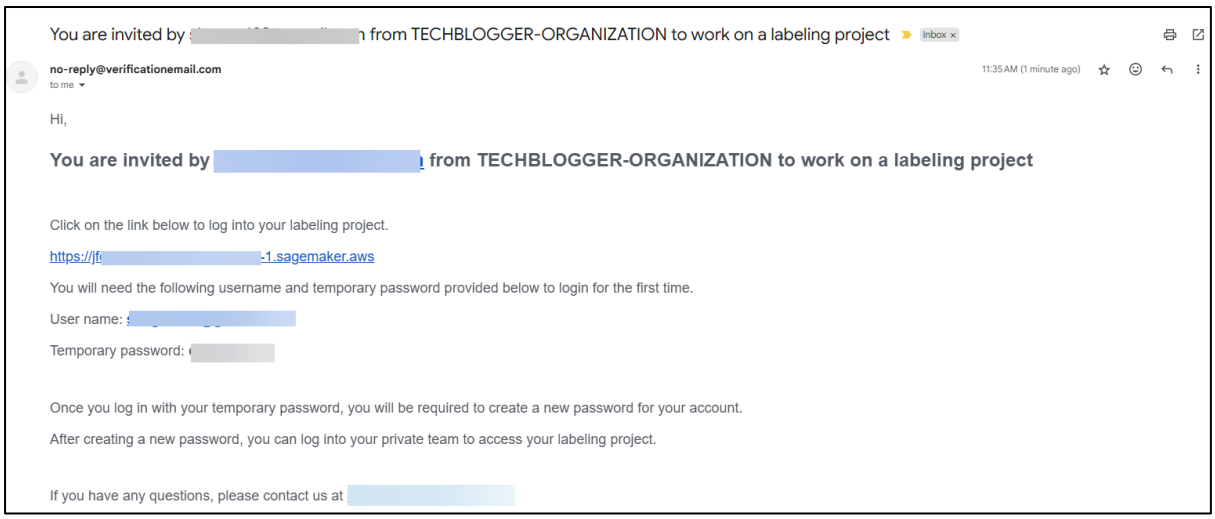
- On the private team details page, go to the **Workers** tab.
- Click **Add workers to team**, select users, and add them.
- You can add more workers later or manage via the Amazon Cognito console.

Tips:

- If you delete all teams, you'll need to recreate the workforce.
- For larger teams, import an existing Cognito user pool or use OIDC (advanced; see docs).

Once created, your workforce is ready. Workers access tasks via the portal link in their invitation email.

The workforce team should be getting an e-mail like the following, They should be able to log in using the provided URL and username/Password in the mail.



Part 2: Creating a Labeling Job (Labeling Your Dataset)

A labeling job applies labels to your dataset using the workforce. The output is a labeled dataset in S3, ready for model training.

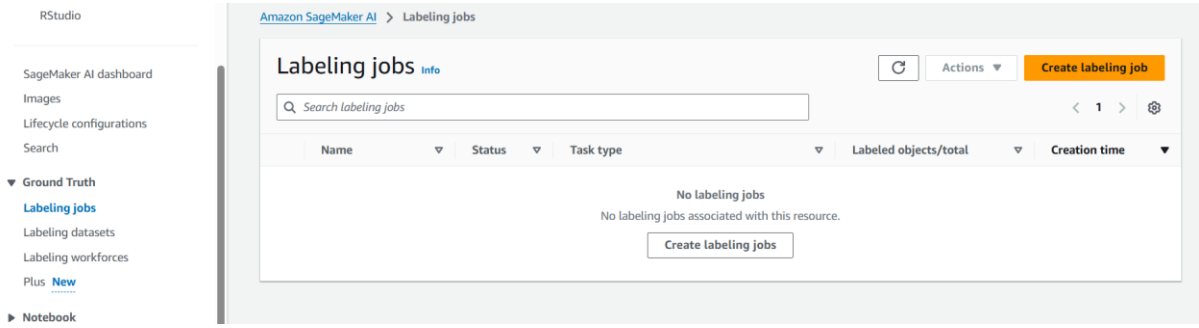
For this example, we'll label text data (e.g., classify numbers as "even" or "odd"). Upload sample files to your input S3 bucket:

- 1.txt with content: "42"
- 2.txt with content: "19"
- 3.txt with content: "21"

Ensure your S3 bucket has CORS configured if using images (not needed for text).

Step 1: Start a New Labeling Job

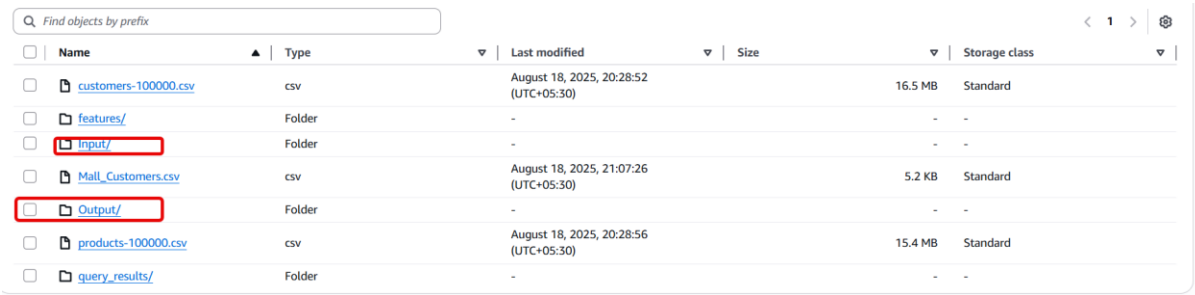
- In the SageMaker console, under **Ground Truth** in the left navigation, select **Labeling jobs**.
- Click **Create labeling job**.



Step 2: Configure Job Overview

- Enter a unique **Job name** (e.g., "text-classification-demo").
- For **Input data setup**, select **Automated data setup**.
- **S3 location for input datasets**: Browse and select your input bucket (e.g., s3://your-input-bucket/).
- **S3 location for output datasets**: Select "Specify a new location" and browse to your output bucket (e.g., s3://your-output-bucket/).
- **Data type**: Choose **Text** (or Image/Video/Object for other types).
- **IAM role**: Select an existing role with SageMaker access or create a new one (allow "Any S3 bucket" for simplicity).

Click **Complete data setup**.



In the S3 bucket, upload all your text files required for the job processing into the “Input” Folder. And create an “Output” folder separately so that once the workforce finished the job execution the output will be stored in this folder

Please Note

For Amazon SageMaker Ground Truth, a manifest file is a JSON Lines (.jsonl) file, where each line describes one object in S3.

My Input files are in s3://custom-sagemaker-bucket-s3-feature-engineering123/Input/

and filenames are 1.txt, 2.txt, 3.txt, 4.txt, the manifest file would look like this:

{"source": "s3://custom-sagemaker-bucket-s3-feature-engineering123/Input/1.txt"}

{"source": "s3://custom-sagemaker-bucket-s3-feature-engineering123/Input/2.txt"}

{"source": "s3://custom-sagemaker-bucket-s3-feature-engineering123/Input/3.txt"}

{"source": "s3://custom-sagemaker-bucket-s3-feature-engineering123/Input/4.txt"}

Steps to use it:

1. Save the above lines in a file called manifest.jsonl (not .json).
2. Upload it to your S3 bucket (e.g., s3://custom-sagemaker-bucket-s3-feature-engineering123/manifest/manifest.jsonl).
3. In **SageMaker Ground Truth**, specify the location of this manifest file when creating a labeling job.

Job overview

Job name

text-analysis-job-creation

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

☐ I want to specify a label attribute name different from the labeling job name.

Label attribute name is the key where your labels are stored in the augmented manifest. Ground Truth uses the labeling job name as the default label attribute name.

Input data setup Info

Use the automated setup to have Ground Truth automatically identify your dataset in S3. Use the manual setup if you have an input manifest file.

○ Automated data setup
Provide the S3 location of the dataset you want labeled and let Ground Truth automatically connect to and use this dataset for your job.

◉ Manual data setup
Provide the S3 location of a file (an input manifest file) that identifies the data objects you want labeled

Input dataset location Info

Provide a path to the S3 location where your manifest file is stored.

s3://custom-sagemaker-bucket-s3-feature- | View | Browse S3

Output dataset location Info

Provide a path to the S3 location where you want your labeled dataset to be stored.

s3://custom-sagemaker-bucket-s3-feature- | View | Browse S3

IAM Role Info

Provide the ID or ARN for your own AWS KMS encryption key for Amazon SageMaker to access your S3 bucket. Choose a role or let us create a role with the AmazonSageMakerFullAccess IAM policy attached.

AmazonSageMaker-ExecutionRole-20250818T084167

▶ Additional configuration - *optional*

Dataset object selection, encryption

Task type Info

Task category

Select the type of data being labeled to view available task templates for it or select 'Custom' to create your own.

Text

Task selection

Select the task that a human worker will perform to label objects in your dataset.

○ Text Classification (Single Label)
Get workers to categorize text into individual classes. Info

○ Text Classification (Multi-label)
Get workers to categorize text into one or more classes. Info

○ Named entity recognition
Get workers to apply labels to words or phrases within a larger text. Info

New Generative AI Tasks Available

New text-based task types are available on SageMaker Ground Truth for generative AI use cases. Click Generative AI on the task category dropdown for access to the text ranking and question and answer task types.

▶ Tags - *optional*

Cancel | Next

Step 3: Select Task Type

- Under **Task type**, choose a category (e.g., **Text**).
- For **Task selection**, pick **Text classification (single label)**.
- Click **Next**.

Step 4: Configure Workers and Tool

- Under **Workers**, select **Private**.
- Choose your private team from the **Private teams** dropdown.
- Set **Task timeout** (e.g., 1 hour) and **Task expiration time** (e.g., 10 days).
- In the labeling tool section:
    - **Task description**: Enter brief instructions, e.g., "Classify the number as even or odd."
    - **Labels**: Add categories, e.g., "Even" and "Odd."
    - **Short instructions**: Provide examples, e.g., "Select 'Even' if divisible by 2."
- Click **Preview** to see what workers will view.

Select workers and configure tool

All fields are required unless otherwise specified

Workers Info

Worker types

- ○ **Amazon Mechanical Turk** — An on-demand 24/7 workforce of over 500,000 independent contractors worldwide powered by Amazon Mechanical Turk.
- ◉ **Private** — A team of workers that you have sourced yourself, including your own employees or contractors for handling data that needs to stay within your organization.
- ○ **Vendor managed** — A curated list of third party vendors that specialize in providing data labeling services, available via the AWS Marketplace.

Private teams

Choose from the teams you created in the private workforce or if you need to create a new team, save your progress and go to Labeling workforces to create a new one.

Image-Classification-Testing-Team

Task timeout

The maximum time a worker can work in a single task. Please see here for information on default and maximum values.

0 hours 5 mins 0 secs

Task expiration time

The amount of time that a task remains available to workers before expiring. Please see here for information on default and maximum values.

10 days 0 hours 0 mins

0 secs

☐ Enable automated data labeling Info

Amazon SageMaker will automatically label a portion of your dataset. It will train a model in your AWS account using Built-in Algorithm and your dataset. When you enable this, training jobs use new computing resources on your behalf. For cost information, See SageMaker pricing

▶ **Additional configuration - *optional***

Workers per dataset object

Image classification (Single Label) labeling tool

Preview

Provide labeling instructions with examples below for workers. Workers will be viewing these instructions when they perform your task. Workers can choose up to 30 labels. See guidelines for See guidelines for creating high-quality instructions

H1 H2 B I A

Good example

Enter description to explain the correct label to the workers

Add image here

Bad example

Enter description of an incorrect label

Add image here

Enter a brief description of the task

The expected output would be to notice if a number is odd or even.

s3://custom-sagemaker-bucket-s3-feature-engineering123/Input/4.txt

PREV NEXT

Select an option

Add up to 30 labels

Odd Number

Even Number

Add new label

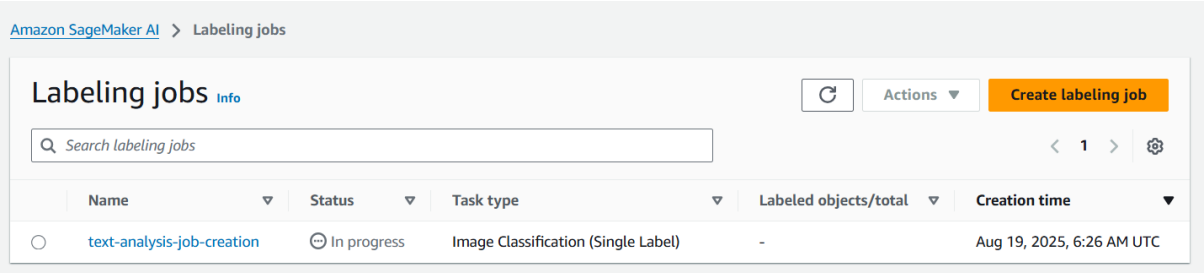
You can add 28 more labels.

▶ Additional instructions - *optional*

Cancel Previous Create

Step 5: Create and Monitor the Job

- Click **Create**.
- The job status will show as "In progress." Workers receive notifications (if SNS set up) and can log in to the portal to label.
- Monitor progress in the Labeling jobs page.
- Once complete, labeled data (manifest file with annotations) is saved to your output S3 bucket.



Part 3: Worker Perspective (Performing Labeling)

- Workers log in via the portal link from their email.
- Select the job and click **Start working**.
- Label each item (e.g., read the text and select "Even" or "Odd").
- Submit when done.

Results appear in the output manifest file, e.g., JSON lines with source data and labels.

Conclusion

You've now created a private workforce and launched a labeling job to produce a labeled dataset! This output can train SageMaker models or other ML workflows. For advanced features like custom templates or auto-labeling, explore the docs.

If you encounter issues, check IAM permissions or console logs. Scale up by adding more workers or using vendor workforces.