Analyze and Prepare Data with Amazon SageMaker Data Wrangler and Amazon EMR

**SPL-TF-300-MLDPML-1 - Version 1.0.0**

© 2024 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. All trademarks are the property of their owners.

Note: Do not include any personal, identifying, or confidential information into the lab environment. Information entered may be visible to others.

Corrections, feedback, or other questions? Contact us at [*AWS Training and Certification*](https://support.aws.amazon.com/#/contacts/aws-training).

**Lab overview**

A local not-for-profit citizen advocacy group promotes government assistance services to qualified citizens in their region. The advocacy group is preparing to promote a newly established assistance service. One of the key eligibility criteria is that an individual cannot have an annual income that exceeds 50,000 USD. The citizen advocacy group has demographic data that includes citizens’ education and employment, but only a subset of the data includes income data. The citizen advocacy group has asked AnyCompany Consulting to deliver a machine learning (ML) solution that can predict if an individual is likely to make less than 50,000 USD, based on the other demographic data. This solution will help them to make the best use of their limited budget by targeting promotions to citizens who are most likely to be eligible for the service. AnyCompany Consulting wants you to visualize and prepare the data. After the data is cleaned, they want you to export the data to their Amazon Simple Storage Service (Amazon S3) bucket.

In addition to that, you have some datasets that AnyCompany Consulting wants you to bring into Amazon SageMaker later that require data preparation at scale. You want to test how Apache Spark on Amazon EMR can be integrated with Amazon SageMaker Studio to process data and prepare it.

In this lab, you learn how to visualize and prepare data, and complete several transformations on the dataset in SageMaker Data Wrangler. After the data is transformed, you learn to export the data back to Amazon S3. You also learn how to discover and securely connect to an EMR cluster directly from SageMaker Studio. You use a SageMaker Studio notebook to visually discover, authenticate with, and connect to an EMR cluster. You then query an Apache Hive table on Amazon EMR using Apache Spark.

OBJECTIVES

After completing this lab, you should be able to do the following:

* Choose effective methods for visualizing data
* Explain the value of data cleaning and transformation
* Describe how to process missing values, outliers, duplicated data, etc.
* Define Key Encoding techniques
* Describe how to ingest and transform data into Amazon Sagemaker Data Wrangler
* Describe how to transform data using Spark on Amazon EMR

TECHNICAL KNOWLEDGE PREREQUISITES

To successfully complete this lab, you should have knowledge of:

* Basic navigation of the AWS Management Console.
* An understanding of database concepts, MySQL, and database availability.

ICON KEY

Various icons are used throughout this lab to call attention to different types of instructions and notes. The following list explains the purpose for each icon:

* **Caution:** Information of special interest or importance (not so important to cause problems with the equipment or data if you miss it, but it could result in the need to repeat certain steps).
* **Learn more:** Where to find more information.
* **Note:** A hint, tip, or important guidance.
* **Task complete:** A conclusion or summary point in the lab.

**Start lab**

1. To launch the lab, at the top of the page, choose **Start lab**.

**Caution:** You must wait for the provisioned AWS services to be ready before you can continue.

1. To open the lab, choose **Open Console**.

You are automatically signed in to the AWS Management Console in a new web browser tab.

**WARNING:** **Do not change the Region unless instructed.**

COMMON SIGN-IN ERRORS

**Error: You must first sign out**



If you see the message, **You must first log out before logging into a different AWS account:**

* Choose the **click here** link.
* Close your **Amazon Web Services Sign In** web browser tab and return to your initial lab page.
* Choose **Open Console** again.

**Error: Choosing Start Lab has no effect**

In some cases, certain pop-up or script blocker web browser extensions might prevent the **Start Lab** button from working as intended. If you experience an issue starting the lab:

* Add the lab domain name to your pop-up or script blocker’s allow list or turn it off.
* Refresh the page and try again.

LAB ENVIRONMENT

The basic architecture of this lab is that two similar datasets have been loaded into an Amazon S3 bucket. An Amazon SageMaker Studio instance with access to those datasets has been created for your use, as well as an Amazon EMR cluster.

AWS SERVICES NOT USED IN THIS LAB

AWS service capabilities used in this lab are limited to what the lab requires. Expect errors when accessing other services or performing actions beyond those provided in this lab guide.

**Task 1: Import, visualize, and perform a preliminary analysis of the data with SageMaker Data Wrangler**

Before the demographic data can be transformed and used effectively to build an ML model, it must be analyzed. AnyCompany Consulting wants you to take the demographic data, import it into SageMaker Data Wrangler and analyze the data. Based on the findings from your analysis in this task, you transform the data in the next task.

SageMaker Data Wrangler is a SageMaker Studio feature for exploration and transformation of image data and tabular data for ML use cases without coding. It includes built-in data analysis capabilities for charts and time-saving model analysis capabilities such as feature importance, target leakage, and model explainability.

In this task, you import and analyze data with a graphical interface by using a notebook and SageMaker Data Wrangler.

**Learn more:** Refer to *Prepare ML Data with Amazon SageMaker Data Wrangler* in the *Additional resources* section for more information about Amazon SageMaker Data Wrangler.

TASK 1.1: ACCESS AND CREATE A SAGEMAKER DATA WRANGLER DATA FLOW IN SAGEMAKER STUDIO

A SageMaker Data Wrangler flow is a series of data preparation steps that you perform on your data. Each transformation is done using a transform step. The flow has a series of nodes that represent importing your data, analyzing your data, and transforming your data.

In this task, you access and create a new SageMaker Data Wrangler flow in SageMaker Studio.

1. Copy the **SageMakerStudioUrl** value that is listed to the left of these instructions.

**Caution:** This lab requires the **Chrome** or **Safari** browsers. It does not work with **Firefox**.

1. Open a new browser tab, and paste the **SageMakerStudioUrl** into the address bar.
2. Press **Enter**.
3. Wait for SageMaker Studio to launch.

**Note:** If you see a pop-up window asking if you want to take a tour of Sagemaker Studio, choose to **Skip Tour for now**.

1. In the left navigation section, locate the **Applications** pane.
2. From the Applications pane, choose **Studio Classic**.

The browser displays the **SageMaker Studio Classic** page.

1. For the listed **Studio Classic** application, in the **Action** column, choose **Open**.

A new browser tab opens, taking you to the **Amazon SageMaker Studio Classic** interface.

**Note:** It might take 1–2 minutes for the SageMaker Studio Classic interface to load for the first time.

**Note:** If you see a *The loading screen is taking a long time. Would you like to clear the workspace or keep waiting?* message, choose Keep waiting.

1. On the left side of SageMaker Studio, expand the **Data** section and choose **Data Wrangler**.

SageMaker Studio opens the **Data Wrangler** tab.

1. Choose **+** **Create Data Wrangler flow**.

**Note:** SageMaker Studio opens the **untitled.flow** tab.

**Note:** A flow file can take 2–3 minutes to set up.

**Note:** When creating a new flow file, you might get an error message that says ***The following instance type is not available: ml.m5.4xlarge. Try selecting a different instance below.*** If so, you can either choose another instance type or close the **untitled.flow** tab and reopen the **untitled.flow** file from the file browser. If you try another instance type, try **ml.m5.8xlarge** first.

You have accessed and created a SageMaker Data Wrangler data flow in SageMaker Studio.

TASK 1.2: LOAD THE DATA FROM AMAZON S3 INTO SAGEMAKER DATA WRANGLER

SageMaker Studio opens the **Create connection** page inside of the **Data Wrangler** tab.

1. Open the context (right-click) menu on the **untitled.flow** file tab, and then, to change the file name, choose **Rename Data Wrangler Flow…**.

SageMaker Studio opens the **Rename File** message window.

1. For **New Name**, enter

DataWranglerLab.flow

.

1. Choose Rename .

The **Rename File** message window closes.

Next, you import the dataset into SageMaker Data Wrangler from Amazon S3. The dataset includes numerical and categorical features of adults, and your task is to develop a model that predicts if their income is below 50,000 USD.

**Note:** SageMaker Data Wrangler also supports importing data from data sources such as Amazon Athena, Amazon Redshift, Snowflake, and Databricks.

**Learn more:** Refer to *Import Data from Amazon S3* in the *Additional resources* section for more information about how to import data into SageMaker Data Wrangler.

1. On the **DataWranglerLab.flow** tab, in the **Available** section, choose **Amazon S3**.

SageMaker Studio opens the **Import tabular, image or time-series data from S3** page inside of the **DataWranglerLab.flow** tab.

1. In the list of buckets, choose (double-click) the Amazon S3 bucket that matches the **LabDataBucket** value that is listed to the left of these instructions.
2. Choose (double-click) the **scripts** folder, choose (double-click) the **data** folder, and then select the **raw\_data.csv** file.
3. Choose Import .

You have imported data from an Amazon S3 source into SageMaker Data Wrangler.

TASK 1.3: GET INSIGHTS ON THE DATA STRUCTURE AND QUALITY

SageMaker Data Wrangler can provide insights and in-depth reports to quickly analyze your datasets.

In this task, you use SageMaker Data Wrangler to generate an initial analysis of the data.

1. On the **DataWranglerLab.flow** tab, choose the **Analysis** tab.

You can use a *Data Quality and Insights Report* to perform an analysis of the data that you imported into SageMaker Data Wrangler. This report helps you know the best ways to clean and process your data.

1. In the **Create analysis** section, configure the following settings:

* For **Analysis type**, select **Data Quality And Insights Report**.
* For **Analysis name**, enter

InitialIncomeReport

.

* For **Target column**, select **income**.
* For **Problem type**, choose **Classification**.

**Note:** You might need to scroll down in the list to locate some parameters.

1. Choose Create .

A summary from the analysis of the data is displayed after a brief wait.

Take some time to review the following sections of the **Data Quality And Insights Report**:

* **Summary**: This section shows important dataset statistics, including the number of features (columns), the percent of data that is missing, the number of duplicate rows, and the different data types identified for each feature.
* **Duplicate rows**: This section shows many of the records that are likely duplicate data. If rows show up here, you might want to add a transformation that removes duplicate data.
* **Anomalous samples**: This section shows records that are anomalies in the dataset. SageMaker Data Wrangler models the data and shows records that have a high chance of being outliers.
* **Target column**: This section shows the distribution of the target column. If a target column is not equally distributed, you need to pay more attention to the records in each value to make sure you are not introducing bias into the model. If the target column is too unbalanced, you can add a transformation to balance the target column. For this lab, the target column is balanced enough to make an effective model.
* **Quick model**: This section shows metrics that are returned from a quick model of the data. You can use quick models to show how effective your transformations are at improving your model performance.
* **Feature summary**: This section shows the prediction power of each feature. Notice that *fnlwgt* and *native\_country* have the lowest prediction power. You can add a transformation to remove these columns. This helps your model focus on the features that have a higher prediction power.
* **Feature details**: This section shows tables of metrics and graphs of each feature. Notice that these metrics and histograms are similar to the graphs that you viewed in the notebook.
* **Definitions**: This section includes definitions of most of the ML terms used in the Data Quality and Insights Report.

Information generated from this report helps guide the transformations in the subsequent steps

**Task complete:** You have used SageMaker Data Wrangler to generate an initial analysis of the data.

**Task 2: Analyze and visualize the data**

You have used SageMaker Data Wrangler to perform a quick visualization and analysis of your data. Now, you want to study your target column to see if there are any additional problems with *income* that you want to address in your dataset. Specifically, you want to see if there are any problems with target leakage and how the target column is distributed across different columns in the dataset.

In this task, you perform an analysis of the currently untransformed dataset using the **Target Leakage** and **Histogram** analysis types before applying any transformations.

**Learn more:** Refer to *Analyze and Visualize* in the *Additional resources* section for more information about SageMaker Data Wrangler analysis types.

TASK 2.1: TARGET LEAKAGE

In this task, you use SageMaker Data Wrangler to check for target leakage in the data.

Target leakage occurs when there is data in an ML training dataset that is strongly correlated with the target label but is not available in real-world data. For example, you might have a column in your dataset that serves as a proxy for the column that you want to predict with your model. SageMaker Data Wrangler calculates the predictive metric of receiver operating characteristic (ROC), which is computed individually for each column through cross validation to generate the target leakage report.

1. To return to the **Data flow** view, choose **< Data flow** located at the top left of the **DataWranglerLab.flow** tab.
2. Choose the **+** icon next to the **Data types** icon, and from the context menu, choose **Add analysis**.
3. In the **Create analysis** section, configure the following settings:

* For **Analysis type**, select **Target Leakage**.
* For **Analysis name**, enter

Income-Target-Leakage

.

* For **Max features**, enter

15

.

* For **Problem Type**, select **classification**.
* For **Target**, select **income**.

1. To generate the **Target Leakage** analysis report, choose Run .

The **Data types · Transform: raw\_data.csv** page is displayed on the **DataWranglerLab.flow** tab.

The report shows that there is no indication of target leakage in your input dataset. However, **native\_country**, and **fnlwgt** have 0.5 predictive ability of ROC. This ROC indicates that these features do not independently provide useful information toward predicting the target. Dropping uninformative columns from the dataset can increase the efficiency of your model training. Before making the decision to drop these features, you should consider whether these could add value when used in tandem with other features. If you want to create a new feature combination, you can create a custom formula or custom transform in SageMaker Data Wrangler. The **fnlwgt** and **native\_country** columns are not useful for training your model. You can drop these columns from the dataset when you are adding your transformations.

You have used SageMaker Data Wrangler to account for target leakage in the data.

TASK 2.2: GENERATING A HISTOGRAM

In this task, you use a histogram to gain insights into target label patterns inside the input dataset.

1. Choose **< All analyses** located above **Edit analysis** section.
2. On the **Analysis** tab, choose Create new analysis .
3. In the **Create analysis** section:

* For **Analysis type**, select **Histogram**.
* For **Analysis name**, enter

Target-Label-Classification

.

* For **X axis**, select **income**.
* For **Color by**, select **race**.
* For **Facet by**, select **sex**.

**Note:** This means that you want to plot histograms by **income** with the **race** factor reflected by a color legend and also faceted by **sex**.

1. To generate the following **Histogram** analysis report, choose Run.

Histograms of the target column for sex and race are shown.

**Task complete:** You performed an analysis of the currently untransformed dataset using the **Target Leakage** and **Histogram** analysis types before applying any transformations.

**Task 3: Perform data transformations and export the datasets**

In the previous tasks, you visualized and analyzed the dataset. In your analysis, you found that there are several transformations that have to be conducted. In this task, you perform data transformations that address issues that you discovered in the dataset.

In this task, you use SageMaker Data Wrangler transformations to prepare the dataset for training.

**Learn more:** Refer to *Transform Data* in the *Additional resources* section for more information about SageMaker Data Wrangler data transforms.

TASK 3.1: DROP COLUMNS AND HANDLE MISSING AND DUPLICATE VALUES

By using the target leakage analysis, you found that **fnlwgt** and **native\_country** do not have strong predictive power, so you can drop the columns. When you looked at the histograms in the notebook, you saw that **occupation** and **workclass** are missing values. The *Data Insights and Quality Report* showed that there are duplicate records in your dataset.

In this task, you drop the unnecessary **fnlwgt**, and **native\_country** columns, remove rows with missing values, remove duplicate records, and clean any string values that might contain spaces to the right or left of the values.

1. To return to the **Data flow** view, choose **< Data flow** located at the top left of the **DataWranglerLab.flow** tab.
2. Choose the **+** icon next to the **Data types** icon, and from the context menu, choose **Add transform**.

First, drop the unnecessary columns that have low prediction power.

1. To list the available transforms, in the **ALL STEPS** section, choose + Add step .
2. In the **ADD TRANSFORM** section, search for and choose

Manage columns

.

1. In the **Manage columns** section, configure the following settings:

* For **Transform**, select **Drop column**.
* For **Columns to drop**, choose the following:
  + **fnlwgt**
  + **native\_country**

1. To drop the columns, choose Preview .
2. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add .

Next, remove any records that have missing values.

1. To list the available transforms, in the **ALL STEPS** section, choose + Add step .
2. In the **ADD TRANSFORM** section, search for and choose

Handle missing

.

1. In the **Handle missing** section, configure the following settings:

* For **Transform**, select **Drop missing**.
* For **Input columns**, choose the following:
  + **occupation**
  + **workclass**

1. To apply the transform, choose Preview .
2. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add .

Then drop the duplicate information.

1. To list the available transforms, in the **ALL STEPS** section, choose + Add step .
2. In the **ADD TRANSFORM** section, search for and choose

Manage rows

.

1. In the **Manage rows** section, for **Transform**, select **Drop duplicates**.
2. To apply the transform, choose Preview .
3. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add .

Finally, to correct any string value errors, remove leading and trailing spaces from the values in categorical columns.

1. To list the available transforms, in the **ALL STEPS** section, choose + Add step .
2. In the **ADD TRANSFORM** section, search for and choose

Format string

.

1. In the **Format string** section, configure the following settings:

* For **Transform**, select **Strip left and right**.
* For **Input columns**, choose the following:
  + **education**
  + **income**
  + **marital\_status**
  + **occupation**
  + **race**
  + **relationship**
  + **sex**
  + **workclass**
* For **Characters to remove**, select the field and press the **space bar** key once. This removes the spaces on both the left and right side of the string.

1. To apply the transform, choose Preview .
2. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add .

You have used SageMaker Data Wrangler to transform and drop unnecessary columns from the dataset.

TASK 3.2: TRANSFORM THE TARGET COLUMN

The goal for the model is to determine whether an individual is likely to make less than 50,000 USD. The target column already includes this information in a string format. To prepare the data for binary classification model training, you need to convert these string values into binary numerical format. In the next steps, you will replace the string values ‘<=50K’ with the numerical value 1, and the string values ‘>50K’ with the numerical value 0.

In this task, you transform data in the target column, **income**.

1. To list the available transforms, in the **ALL STEPS** section, choose + Add step .
2. In the **ADD TRANSFORM** section, search for and choose

Search and edit

.

1. In the **Search and edit** section, configure the following settings:

* For **Transform**, select **Find and replace substring**.
* For **Input column**, choose **income**.
* For **Pattern**, enter

<=50K

.

* For **Replacement string**, enter

1

.

1. To apply the transform, choose Preview .

All **<=50K** values in the target column have been replaced with **1**.

1. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add .
2. To list the available transforms, in the **ALL STEPS** section, choose + Add step .
3. In the **ADD TRANSFORM** section, search for and choose

Search and edit

.

1. In the **Search and edit** section, configure the following settings:

* For **Transform**, select **Find and replace substring**.
* For **Input column**, choose **income**.
* For **Pattern**, enter

>50K

.

* For **Replacement string**, enter

0

.

1. To apply the transform, choose Preview .

All **>50K** values in the target column have been replaced with **0**.

1. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add .

You have used SageMaker Data Wrangler to transform the **income** target column.

TASK 3.3: HANDLE OUTLIERS

When you viewed the histograms in the notebook, you saw that there are outliers for **capital\_gain**. Remove any values greater than 80000 and any values less than 0.

In this task, you handle the outliers in the data.

1. To list the available transforms, in the **ALL STEPS** section, choose + Add step .
2. In the **ADD TRANSFORM** section, search for and choose

Handle outliers

.

1. In the **Handle outliers** section, configure the following settings:

* For **Transform**, select **Min-max numeric outliers**.
* For **Input columns**, choose **capital\_gain**.
* For **Fix Method**, select **Remove**.
* For **Upper threshold**, enter

80000

.

* For **Lower threshold**, enter

0

.

1. To apply the transform, choose Preview .
2. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add .

You have used SageMaker Data Wrangler to handle outliers in the grouped data.

TASK 3.4: PERFORM ORDINAL ENCODING

The **education\_num** column has values that range from 9 to 13. Use ordinal encoding to reset these columns to a range starting at 0. The **education** column has string values, but you want the model to interpret the range of education values. Use ordinal encoding to change the categories to integer values in a range starting at 0.

**Note:** The ordinal encoder turns a categorical feature into a numerical format. The labels are translated to integer values based on their ordinal relationship to each other.

In this task, you use the **encode categorical** built-in transform named **Ordinal encode** to encode categories into an integer between 0 and the total number of categories in the input column that you select.

1. To list the available transforms, in the **ALL STEPS** section, choose + Add step .
2. In the **ADD TRANSFORM** section, search for and choose

Encode categorical

.

1. In the **Encode categorical** section, configure the following settings:

* For **Transform**, select **Ordinal encode**.
* For **Input columns**, choose the following:
  + **education\_num**
  + **education**
  + **occupation**
* For **Invalid handling strategy**, select **Skip**.

1. To apply the transform, choose Preview .
2. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add .

You have used SageMaker Data Wrangler to perform an ordinal encode type of transform to the data.

TASK 3.5: PERFORM ONE-HOT ENCODING

Ordinal encoding can introduce bias in a model if used incorrectly. If a column contains labels that have equal value, like *sex* in this example, you can use one-hot encoding to turn the labels of a column into columns with labels of 0 for false and 1 for true. Transform the **sex** column into two columns, one for male and one for female, using one-hot encoding.

**Note:** The one-hot encoder turns a categorical feature into new columns that correspond with each label in the original column, each with values of 0 and 1.

In this task, you perform **one-hot encoding** to transform the categorical variables into a more useable form for ML.

1. To list the available transforms, choose + Add step .
2. In the **ADD TRANSFORM** section, search for and choose

Encode categorical

.

1. In the **Encode categorical** section, configure the following settings:

* For **Transform**, select **One-hot encode**.
* For **Input columns**, choose the following:
  + **marital\_status**
  + **race**
  + **relationship**
  + **sex**
  + **workclass**
* For **Invalid handling strategy**, select **Skip**.
* For **Output style**, select **Columns**.

1. To apply the transform, choose Preview .
2. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add.

You have used SageMaker Data Wrangler to perform a categorical encode type of transform to the data.

TASK 3.6: MOVE THE TARGET COLUMN

Some algorithms, such as **XGBoost**, require the target column to be the first column.

In this task, you move the **income** column to the first column.

**Learn more:** Refer to *Input/Output Interface for the XGBoost Algorithm* in the *Additional resources* section for more information about XGBoost in SageMaker Studio.

1. To list the available transforms, choose + Add step .
2. In the **ADD TRANSFORM** section, search for and choose

Manage columns

.

1. In the **Manage columns** section, configure the following settings:

* For **Transform**, select **Move column**.
* For **Move type**, select **Move to start**.
* For **Column to move**, select **income**.

1. To apply the transform, choose Preview .
2. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add .

You have used SageMaker Data Wrangler to move the target column.

TASK 3.7: EXPLORE THE DESTINATION AND EXPORT FEATURES

In your SageMaker Data Wrangler flow, you can export some or all of the transformations that you have made to your data processing pipelines. SageMaker Data Wrangler gives you the ability to export your data to a location within an Amazon S3 bucket. You can specify the location using one of the following methods:

* **Destination node** – Where SageMaker Data Wrangler stores the data after it has processed it.
* **Export to** – Exports the data resulting from a transformation to Amazon S3.
* **Export data** – For small datasets, quickly export the data that you have transformed.

In this task, you explore the various sources that SageMaker Data Wrangler can export data to.

1. To return to the **Data flow** view, choose **< Data flow** located at the top left of the **DataWranglerLab.flow** tab.
2. Choose the **+** icon next to the **Move column** icon at the end of your list of transforms.

Several options are available, including **Add destination** and **Export to**.

* With **Add destination**, you can set **Amazon S3** or **SageMaker Feature Store** as destinations for your processed data.
* With **Export to**, you can export processed data to **Amazon S3**, **Amazon SageMaker Pipelines**, or **Amazon SageMaker Feature Store** (among others) by using JupyterNotebook. Or you can export code with **Python Code**, which completes the transformations that you just configured in the SageMaker Studio UI.

In the next task, you split the data into **Training**, **Validation**, and **Test** sets and add a **Destination node** for each data set.

You have explored the various sources that SageMaker Data Wrangler can export data to.

TASK 3.8: EXPORT THE DATA TO AN S3 BUCKET

For this task, you are going to use **Destination node** to specify the location. If you want to output a series of data processing steps that you have performed to Amazon S3, you create a destination node. A destination node tells SageMaker Data Wrangler where to store the data after you have processed it. After you create a destination node, you create an Amazon SageMaker processing job to output the data. When you are using a destination node, it runs the computational resources needed to output the data that you have transformed to Amazon S3.

You can use a destination node to export some or all of the transformations that you have made in your SageMaker Data Wrangler flow. You can use multiple destination nodes to export different transformations or sets of transformations.

Before you can use **Destination node**, you need to create a **Split data transformation** to split datasets within SageMaker Data Wrangler. This transformation splits your dataset into training, test, and optionally validation datasets without you having to write any code.

* **Training** – Used to train an algorithm or ML model. The model iteratively uses the data and learns to provide the desired result.
* **Validation** – Introduces new data to the trained model. You can use a validation set to periodically measure model performance as training is happening, and also tune any hyperparameters of the model. Validation datasets are optional.
* **Test** – Used on the final trained model to assess its performance on unseen data. This helps determine how well the model generalizes.

1. Choose the **+** icon next to the **Move column** icon at the end of your list of transforms, and choose **Add transform**.
2. To list the available transforms, in the **ALL STEPS** section, choose + Add step .
3. In the **ADD TRANSFORM** section, search for and choose

Split data

.

1. For **Transform**, choose the default value **Randomized split**.

**Note:** Random split partitions data randomly into train, test, and, optionally validation datasets by using the percentage specified for each dataset. It ensures that the distribution of the data is similar in all datasets. Choose this option when you do not need to preserve the order of your input data. For example, consider a movie dataset where the dataset is sorted by genre and you are predicting the genre of the movie. A random split on this dataset ensures that the distribution of the data includes all genres in all three datasets.

1. In the **Splits** section, configure the following settings:

* For **Split name**

Train

, set the **Split percentage** at

0.7

.

* For **Split name**

Test

, leave the **Split percentage** at

0.2

.

1. Choose the + icon in the **Split name** section to add an additional split.
2. For **Split name**, enter

Validation

 and for **Split percentage**, enter

0.1

.

**Note:** The split percentage can be any value that you want, provided that all three splits add up to 1 (100%). You can also specify optional fields like Error threshold and Random seed. You can achieve an exact split by setting the error threshold to 0. A smaller error threshold can lead to more processing time for splitting the data. This allows you to control the trade-off between time and accuracy on the operation. For this lab, the value is left at the default setting. The Random seed option is for reproducibility. If this is not specified, SageMaker Data Wrangler uses a default random seed value. You can leave it blank for the purpose of this task.

1. To preview your data split, choose Preview .
2. Choose the dropdown beside **Previewing: Randomized split**, and switch between **Train**, **Test**, and **Validation** to review the details of each split.
3. To save the transform to your **SageMaker Data Wrangler** data flow file, choose Add.

The **Data flow** page appears and shows a new **Randomized split** node with three **Datasets**.

1. On the **Data flow** page, choose Run validation located at the top right corner of the page.
2. Choose the **+** icon next to the first **Dataset: raw\_data.csv (Train)** icon at the end of your list of transforms.
3. Choose **Add destination** and then select **Amazon S3**.
4. In the **Amazon S3** section, configure the following settings:

* For **Dataset name**, enter

adult data train

.

* For **Amazon S3 location**, choose Browse.
* In the list of buckets, open the Amazon S3 bucket that matches the **LabDataBucket** value that is listed to the left of these instructions.
* Navigate to the **scripts/data** subfolder, and select the **train** subfolder.
* Choose Select.
* For **Number of partitions**, choose

1

.

1. Choose Add destination.
2. Choose the **+** icon next to the first **Dataset: raw\_data.csv (Test)** icon at the end of your list of transforms.
3. Choose **Add destination** and then select **Amazon S3**.
4. In the **Amazon S3** section, configure the following settings:

* For **Dataset name**, choose

adult data test

.

* For **Amazon S3 location**, choose Browse .
* In the list of buckets, open the Amazon S3 bucket that matches the **LabDataBucket** value that is listed to the left of these instructions.
* Navigate to the **scripts/data** subfolder, and select the **test** subfolder.
* Choose Select.
* For **Number of partitions**, choose

1

.

1. Choose Add destination.
2. Choose the **+** icon next to the first **Dataset: raw\_data.csv (Validation)** icon at the end of your list of transforms.
3. Choose **Add destination** and then select **Amazon S3**.
4. In the **Amazon S3** section, configure the following settings:

* For **Dataset name**, choose

adult data validation

.

* For **Amazon S3 location**, choose Browse .
* In the list of buckets, open the Amazon S3 bucket that matches the **LabDataBucket** value that is listed to the left of these instructions.
* Navigate to the **scripts/data** subfolder, and select the **validation** subfolder.
* Choose Select.
* For **Number of partitions**, choose

1

.

1. Choose Add destination.
2. On the **Data flow** page, choose Create job.
3. In the **Create job** section, leave everything at the default values and choose Next, 2. Configure job.
4. Choose Create.
5. You can watch the Processing jobs by choosing the **Processing Job name** that contains **DataWranglerLab-** in its name.

When the processing job is complete, the files are added to the Amazon S3 bucket. You do not need to wait for the processing job to finish.

**Task complete:** You used SageMaker Data Wrangler transformations to prepare the dataset for training. Then you split and exported the data to Amazon S3.

**Task 4: Set up the environment**

In this task, you launch a SageMaker Studio application and access your lab resources.

Next, you clone a Git repository and open your lab repository folder.

1. In the left menu bar, choose the **Git** icon.
2. Choose Clone a Repository.

The SageMaker Studio environment displays the **Clone Git Repository** window.

1. For **Git repository URL (.git):**, copy and paste the **CloneUrlForRepo** value listed to the left of these instructions, and select the suggested url from the dropdown menu.
2. Choose Clone.
3. Wait for the repository cloning to complete. The folder shows in a pane on the left side of the screen.

The SageMaker Studio environment opens the **labxrepo** folder in a pane on the left side of the screen when the clone completes.

**Task complete:** You have successfully launched SageMaker Studio and cloned a Git repository into the environment.

**Task 5: Connect to an EMR cluster**

In this task, you use a SageMaker Studio notebook to visually discover, authenticate with, and connect to an EMR cluster. By establishing this connection, you can use the notebook to interact with the data hosted on Amazon EMR.

1. From the left navigation pane, open the **labnotebook.ipynb** notebook.

The SageMaker Studio environment displays the **Set up notebook environment** window.

1. In the **Set up notebook environment** window, select the following:

* **Image**: SparkAnalytics 2.0
* **Kernel**: SparkMagic PySpark

1. Choose Select.

**Note:** It might take a moment for the kernel to load.

1. In the upper–right area of the notebook, choose **Cluster**.

The SageMaker Studio environment displays the **Connect to cluster** window.

1. Choose the cluster with the name **EMR-Cluster-LabX** and cluster ID starting with **j-**.
2. Choose Connect.

The SageMaker Studio environment displays the **Select credential type for “EMR-Cluster-LabX”** window.

1. In the **Select the credential type for “EMR-Cluster-LabX”** dialog box, choose the **No credential** option.

**Note:** For this lab, for simplicity, you use a No-Auth authentication.

1. Choose Connect.

**Note:** It takes approximately 1–2 minutes to connect to the cluster.

This adds a code block to the active cell in the notebook and runs automatically to establish a connection.

After the Spark application has started, a **SparkSession available as ‘spark’** message is displayed in the notebook.

**Task complete:** You have connected the SageMaker Studio notebook to an Amazon EMR cluster.

**Task 6: Explore and query data from the SparkMagic PySpark kernel**

In this task, you perform exploratory data analysis using Apache Spark on the SparkMagic PySpark kernel.

**Note:** While similar in many respects, using the SparkMagic PySpark kernel is different than using Data Wrangler, and therefore the data transformations and results may not be exactly the same. Both methods have advantages.

1. Carefully advance through the **labnotebook.ipynb** notebook. Run each code cell and review its output. To run a cell, select within the cell and press **Shift + Enter** or, at the top of the page, choose the **Run** button. An asterisk appears next to the code block while it runs.

When you have finished running the notebook, return here to end the lab.

**Task complete:** You have performed data analysis in the notebook using the SparkMagic PySpark kernel.

**Conclusion**

You have successfully done the following:

* Chosen effective methods for visualizing data
* Explained the value of data cleaning and transformation
* Described how to process missing values, outliers, duplicated data, etc.
* Defined Key Encoding techniques
* Described how to ingest and transform data into Amazon Sagemaker Data Wrangler
* Described how to transform data using Spark on Amazon EMR

**End lab**

Follow these steps to close the console and end your lab.

1. Return to the **AWS Management Console**.
2. At the upper-right corner of the page, choose **AWSLabsUser**, and then choose **Sign out**.
3. Choose **End lab** and then confirm that you want to end your lab.

**Additional Resources**

* [Prepare ML Data with Amazon SageMaker Data Wrangler](https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler.html)
* [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/adult)
* [Analyze and Visualize](https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-analyses.html)
* [Import Data from Amazon S3](https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-import.html#data-wrangler-import-s3)
* [Transform Data](https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-transform.html)
* [Similarity Encode](https://docs.aws.amazon.com/sagemaker/latest/dg/data-wrangler-transform.html#data-wrangler-transform-cat-encode-similarity)
* [Input/Output Interface for the XGBoost Algorithm](https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html#InputOutput-XGBoost)

For more information about AWS Training and Certification, see [*https://aws.amazon.com/training/*](https://aws.amazon.com/training/).

*Your feedback is welcome and appreciated.*  
If you would like to share any feedback, suggestions, or corrections, please provide the details in our [*AWS Training and Certification Contact Form*](https://support.aws.amazon.com/#/contacts/aws-training).