README

**Building a LEAF-ESP solution in Microsoft Azure**

*Introduction*

Welcome to the documentation for “Building a LEAF-ESP solution in Microsoft Azure.” The purpose of this repository is to provide information, code, and guidance to assist your team in delivering a data science solution using the LEAF-ESP framework on the Azure cloud, within the Azure Machine Learning service. This repository will center around an example problem and provide an end-to-end solution.

*Section 1: Creating the workspace*

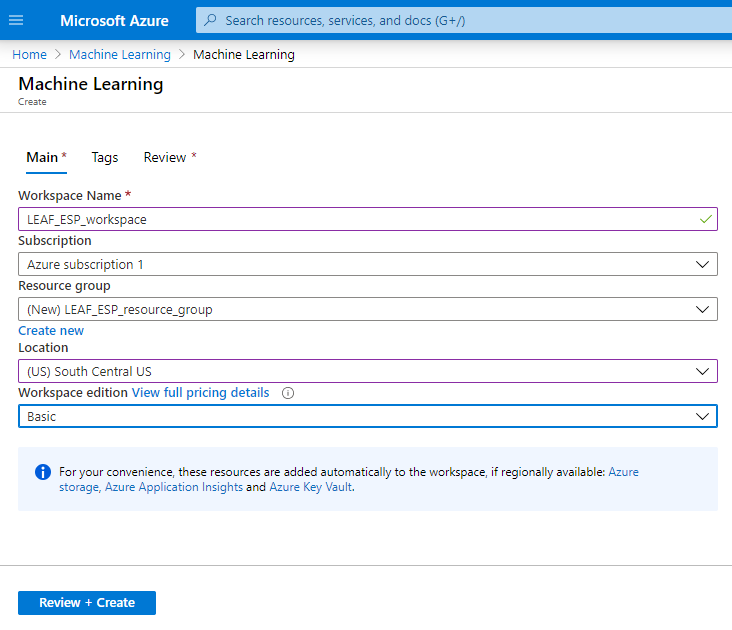
The first aspect of the Azure ML service is the workspace. A workspace is a context for experiments, data storage, compute targets, etc. within a ML workload. Workspaces define the boundary for related ML assets. They can be used to group projects (say you want to build several models using a common data source), deployment environments (a test environment vs a production environment), teams, etc. The assets which can be contained in a workspace include:

* Compute targets for development, training, and deployment.
* Data for experimentation and model training.
* Notebooks containing shared code and documentation.
* Experiments, including run history with logged metrics and outputs.
* Pipelines that define orchestrated multi-step processes.
* Models that you have trained.

When you create a workspace, you are creating a Resource Group. And by default, there are other resources which are created in order to properly interact with the workspace (we will utilize most of these default resources throughout the solution for various purposes):

* A storage account – used to store files and/or data to be used by the workspace
* An Application Insights instance – used to monitor models/services within the workspace
* An Azure Key Vault instance – used to manage authentication keys, credentials, and other secrets to access the workspace
* Virtual Machines – with their associated virtual hardware resources, used for compute jobs
* A container registry – used to manage containers for deployed models

There are multiple ways to create a ML workspace, but the simplest is though the Azure portal (aka the console). Once you create your Azure account, you should navigate to the Machine Learning service from the search bar. From there you should create a new resource, specifying a unique workspace and resource group name while selecting a location close to you. For **Edition** selecting “Enterprise” is optimal for a real-world setting, as it enables the no-code Azure Designer pipeline building service (which allows you to visually create ML pipelines with a drag-and-drop interface while integrating code and/or built-in modules) and graphical tools for auto-ML or data drift monitoring. These may be pretty important omissions depending on your use-case. But for our purposes the solution can be done with the “Basic” edition, the most important consideration being the Region, you must ensure the region allows for GPU processing for the genetic algorithm to run in a timely fashion. See the example below for a configuration.



Once you create your workspace, you can manage all of your resources in the portal. But Microsoft has a UI specifically for data science which is simpler to use for data science tasks called Machine Learning Studio. You can search for this in the top search bar and specify the workspace you just created as the workspace to connect to. From there, you can manage all of your ML assets and easily create new Jupyter Notebooks to use the python SDK.

The next step is to create a compute instance:

1. In the Azure Machine Learning studio web interface for your workspace, view the \*\*Compute\*\* page. This is where you'll manage all the compute targets for your data science activities.

2. On the \*\*Compute Instances\*\* tab, add a new compute instance, giving it a unique name and using the \*\*STANDARD\_DS1\_V2\*\* VM type template. Remember with a computationally intensive algorithm like LEAF, it may be a good idea to select GPU VM type, but this is optional.

3. While the notebook VM is being created, switch to the \*\*Compute Clusters\*\* tab, and add a new training cluster with the following settings:

\* \*\*Compute name\*\*: LEAF-cluster

\* \*\*Virtual Machine size\*\*: Standard\_DS1\_v2

\* \*\*Virtual Machine priority\*\*: Dedicated

\* \*\*Minimum number of nodes\*\*: 0

\* \*\*Maximum number of nodes\*\*: 2

\* \*\*Idle seconds before scale down\*\*: 120

*Section 2: Starting a Notebook and Uploading Data*

Once your compute instances and clusters are up and running, you can now create a Jupyter notebook and use the SDK. Click on the Notebooks section and select “create new file”. Name it something like LEAF-notebook and ensure it is connected to your cluster. Select “Edit” -> “Edit in Jupyter” to connect to your notebook.

We now need to create and register a datastore so we can upload data to a Blob storage account. Go to the portal, then add a new \*\*Storage account\*\* to the resource group with the following settings:

- \*\*Storage account name\*\*: A unique name.

- \*\*Location\*\*: The same location as your workspace

- \*\*Performance\*\*: Standard

- \*\*Account kind\*\*: StorageV2 (general purpose v2)

- \*\*Replication\*\*: Locally-redundant storage (LRS)

- \*\*Access tier (default)\*\*: Hot

- Use the default settings for networking

Wait for the storage account to be created, and then go to the resource in the portal. Open the \*\*Containers\*\* page for the storage account, and add a container with the following settings:

- \*\*Name\*\*: leaf-data

- \*\*Public access level\*\*: Private (no anonymous access)

After your container has been added, view the \*\*Access Keys\*\* page for your storage account and copy \*\*key1\*\* to the clipboard - you will need this in the next task.

Create a new datastore with the following settings:

- \*\*Datastore name\*\*: leaf\_data

- \*\*Datastore type\*\*: Azure Blob Storage

- \*\*Account selection method\*\*: From Azure subscription

- \*\*Subscription ID\*\*: \*Your Azure subscription\*

- \*\*Storage account\*\*: \*The storage account you created in the previous task\*

- \*\*Blob container\*\*: leaf-data

- \*\*Authentication type\*\*: Account key

- \*\*Account key\*\*: \*Paste the key you copied in the previous task\*

After the datastore has been added, verify that it is listed on the \*\*Datastores\*\* page.