Mount an EFS file system on a SageMaker Instance

SageMaker Instances have fixed persistent memory of 5GB, and around 10GB of non-persistent in the /tmp directory, which is cleaned every time the instance is stopped or restarted. Those limits could be troublesome if you need to load or create files larger than that. The solution is to use Amazon EFS storage service and to mount a file system to your instance. This service is more expensive than s3 but faster, it's advantage compare to EBS is it's elasticity. An EFS file system will also be useful to share notebooks between SageMaker users.

Step 1: Create an Elastic File System (EFS)

As this part has already been well explained by Mark and Dan (https://wiki.ucar.edu/display/JEDI/Building+JEDI+on+AWS+with+Singularity) I'm just going to quote them:

"So, to set up an EFS file system you can follow the instructions in this tutorial:

https://aws.amazon.com/getting-started/tutorials/create-network-file-system/?trk=gs_card

In short, it involves going to your AWS console and navigating to Storage - EFS. Select Create File System and then just go through the menus and select the default values for everything. Make sure the VPC in particular is the default VPC (unless you deliberately want to use something else). After you get through all the menus, you'll see a button on the lower right called Create File System.

To see information on your file system at any time, go to the main EFS dashboard, select File systems from the left menu, and select the file system you want."

I have already created a file system for Machine Learning purpose. We have to use the same if we want to share files and notebooks. The one I have created is named "ml_efs_file_system", and it's id is "fs-4a916d01".

Step 2: Create your SageMaker Instance

You can follow Amazon's tutorial:
https://docs.aws.amazon.com/sagemaker/latest/dg/gs-setup-working-env.html
In order to link our efs to the instance choose the default VPC, then select a subnet and defaults security groups (sg-f6c566be).

We can now add a lifecycle configuration in order to automatically mount the file system on our instance every time we start it. I have created one called "efs-mount" but I'm not sure you can access it. So if you can't, choose create a new lifecycle configuration and use this script:

```
#!/bin/bash

set -e

mkdir efs

sudo mount -t nfs \
    -o nfsvers=4.1,rsize=1048576,wsize=1048576,hard,timeo=600,retrans=2 \
    172.31.27.234:/ \
    ./efs

sudo chmod go+rw ./efs
```

Replace the IP address by the one corresponding to the subnet zone you chose previously. To know the zone's IP address you can go to the file system console and click on the one corresponding. This one in the script (172.31.27.234) is for us-east-1a.

You can now click on the create notebook instance button, everything is set!

Miscellaneous

Within your notebooks note that you have to go to the parent directory to access the efs repository. So if you want to read an h5 file for example:

```
import pandas as pd
df_2D = pd.read_hdf("../efs/data/df_2D.h5")
```

I have created two repository at the root of the efs one: data and notebooks. You can open a terminal from your instance if you want to manage it with unix commands.

To share your work you can use the terminal to copy your notebooks in the efs/notebooks directory before closing your instance:

cp SageMaker/*.ipynb efs/notebooks

I know it’s tedious so if someone has another solution for that it would be very helpful.

Useful links for SageMaker:
https://aws.amazon.com/sagemaker/pricing/instance-types/
https://aws.amazon.com/sagemaker/pricing/
https://forums.aws.amazon.com/forum.jspa?forumID=285&start=0

Do not hesitate if you have any questions!