**Operationalizing a Machine Learning Project**

**Notebook instance creation**

We chose a ml.t2.medium instance type. This is because the actual notebook performs light operations and we don't need more power. The delegated instances perform the heavy workload, but we'll see that later.

Texto, Aplicación

Descripción generada automáticamente

Then, we create a new role. We only need the notebook to access to the edgarin-mlend bucket, so we configure the new role that way:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

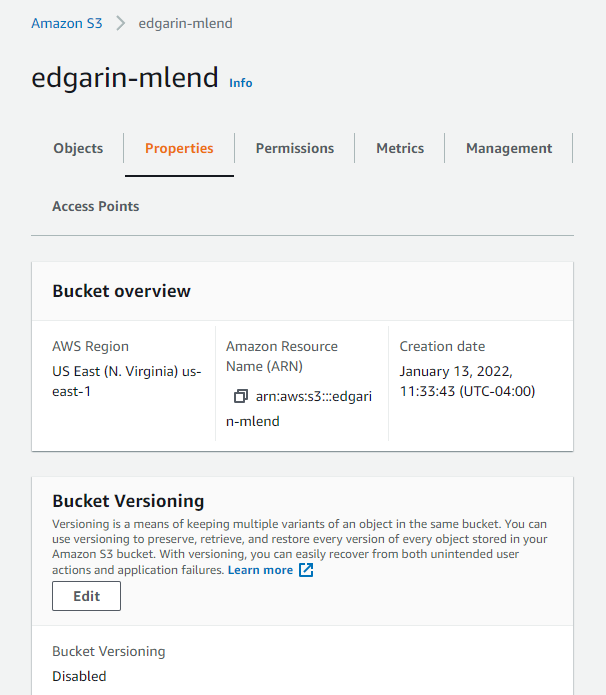
Next step is to configure the other security settings. This includes

* Selecting the newly created role
* Disable root access to the notebook
* Select a specific vpc, subnet and security group

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

We also created a bucket to store the training data and the artifacts (the aforementioned edgarin-mlend bucket).

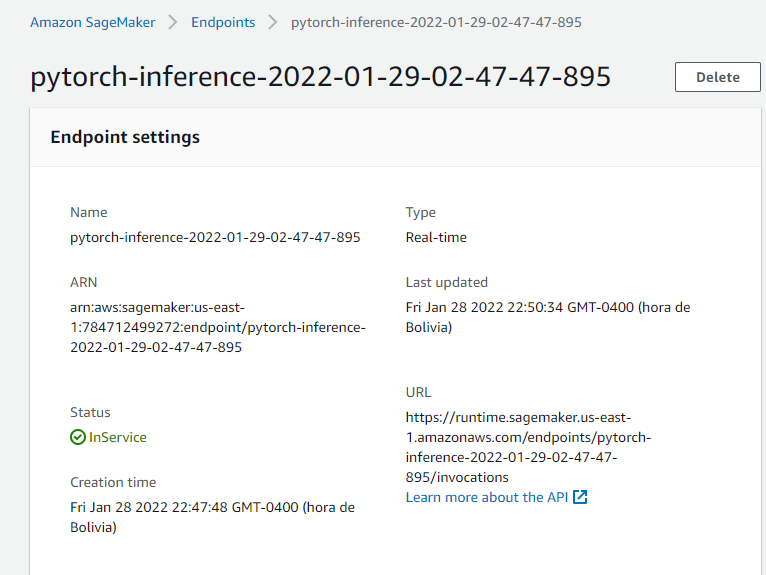


**Training and Endpoint deployment**

All the process is performed in the train-and-deploy.ipynb notebook. In a nutshell, it does the following

* Data setup
* Training and HPO
* Best model deployment

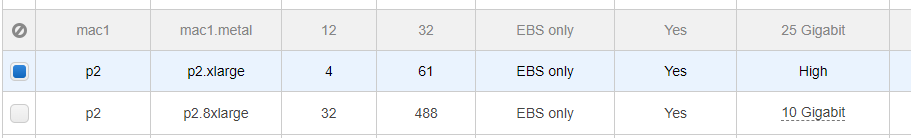
After all this is done, the endpoint is deployed



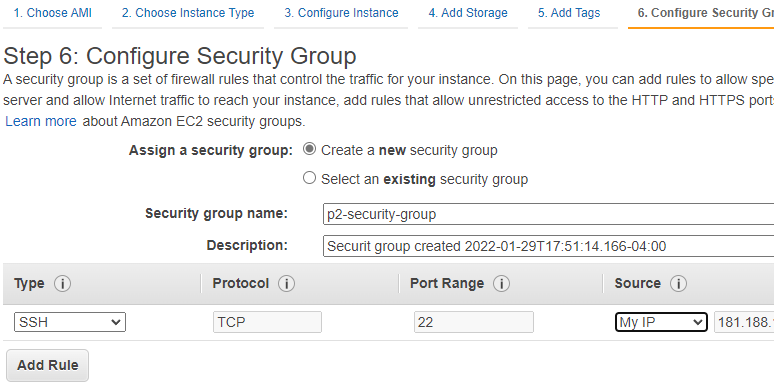
**Training on EC2 instance**

The choice is an ml.m5.xlarge instance to be able to train in a reasonable amount of time without much cost. Given this will be a 1 time run, a spot instance was chosen.

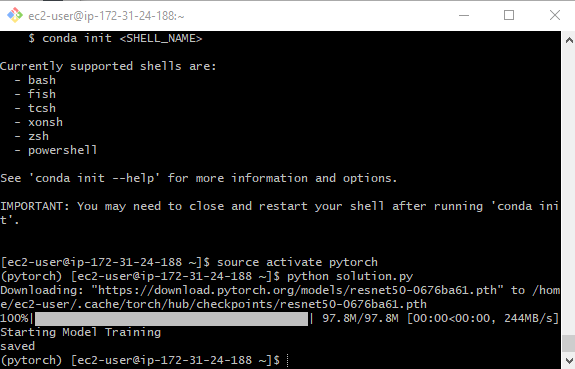
If the code had had CUDA, we would have chosen a p2.xlarge instance (cheapest with good GPU).



For the security group, we chose to access it only from my ip with a keypair (ssh)



Then, training is triggered via ssh  This creates a model artifact in TrainedModels/model.pth



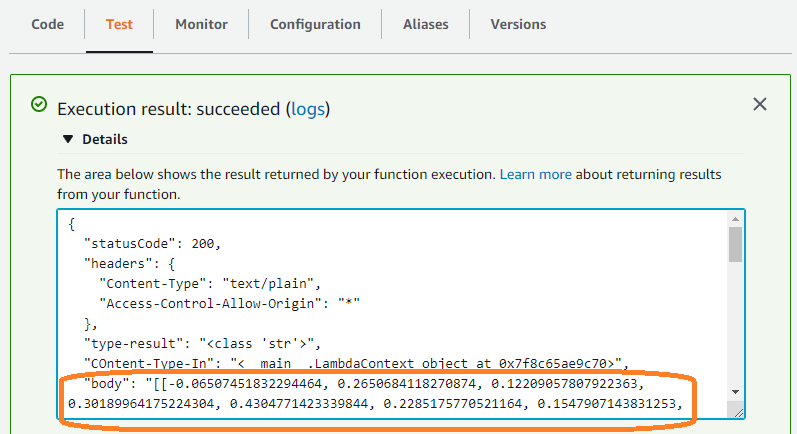
This script (ec2train1.py) is essentially the same as the one trained in script mode (hpo.py) with the following minor differences:

* hpo.py does have logging
* hpo.py is parameterized by command line or env vars (learning rate, batch size, file locations), whereas ec2train1.py has them hardcoded In fact, we could use hpo.py for both purposes, as long as we send the params when calling it.

**Endpoint invocation from Lambda**

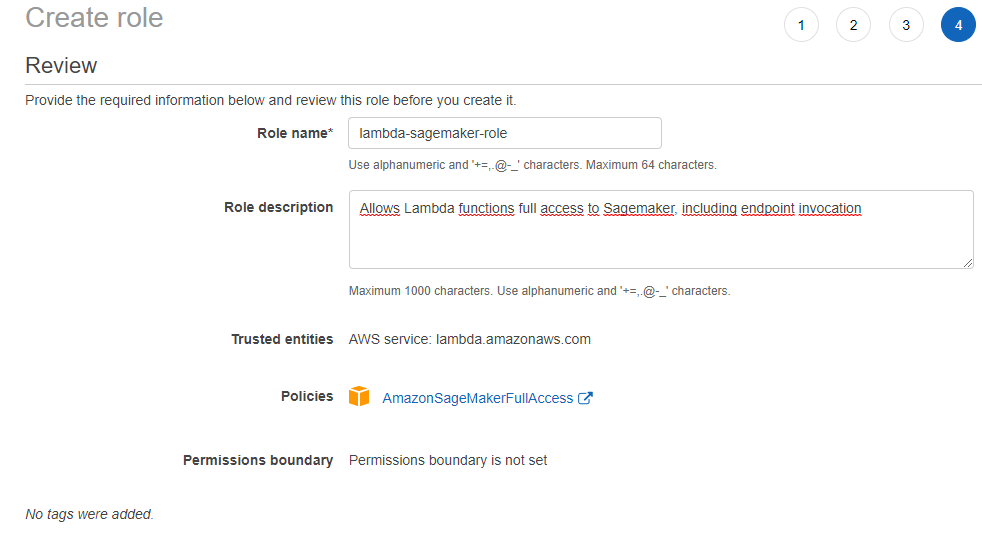
We create a lambda function that will basically take a dog image URL and will invoke the deployed endpoint to perform the prediction. It uses boto3's invoke\_endpoint api, sending a json with the image url. Let's keep in mind that behind the scenes, the inference2.py script will transform this url to a binary image that the endpoint can process.

Finally, the prediction -as defined in the training script- is an array of log probabilities for all breeds:

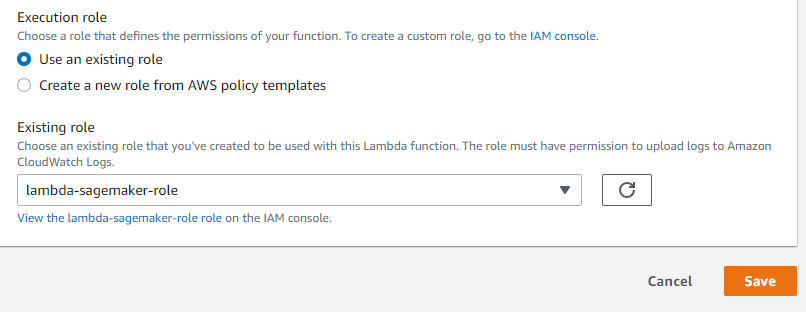


**About Lambda's security**

Before executing the lambda, function, its execution role needs permisions to invoke the endpoint. So, a role with AWSSagemakerFullAccess policy was created:



And then of course the role is attached to the lambda function.



In a real life scenario, it's strongly suggested to apply the least privilege principle. According to this, we would create a role with only permissions to invoke the lambda, but not full sagemaker access.

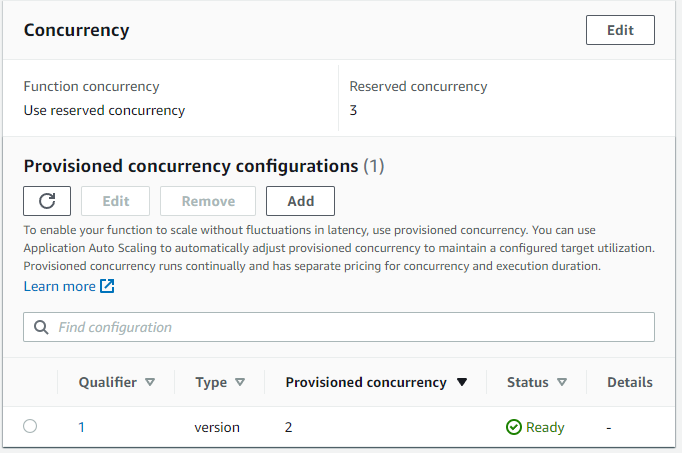
On the other hand, it's very important to configure resource based policies to define who specifically can execute the lambda function. For instance, if it's through API gateway, restrict the permissions to this service (and also configuring the correct authentication/authorization permissions in the api itself, but that's another story).

Finally, adding the lambda to a VPC needs to be considered, if applicable.

**Concurrency**

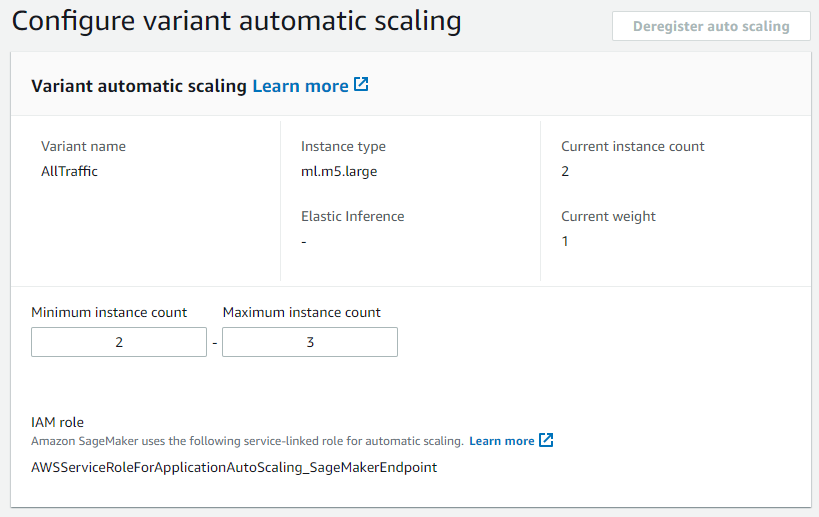
We setup concurrency/autoscaling for both the lambda function and the endpoint

For lambda, we reserved 3 instances and also provisioned concurrency of 2 always-on with the possibility to scale to a 3rd instance.

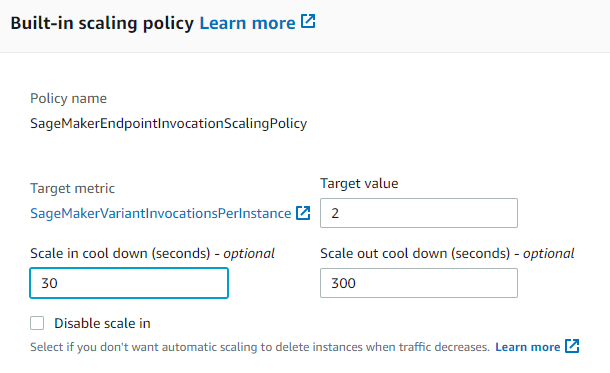


For the endpoint, we configured autoscaling with a minimum of 2 and maximum of 3

instances



Also, we want to spin up instances when there are 5 or more simultaneous requests (Typo: Screenshot shows 2 as target value but the final config was actually 5)



Summarizing, we chose autoscaling of minimum 2 and maximum 3 instances for both lambda and endpoint. This is because due to our synchronous design, we can expect a 1-1 correspondence here. 1 lambda will derive the request to 1 endpoint.

However, if we had an async design, we could have many more lambdas to attend the requests, queue them and then have less endpoints to make the inferences as they get free.