Welcome to the first lab!

We’re going to make some assumptions before we kick things off. We are assuming that you:

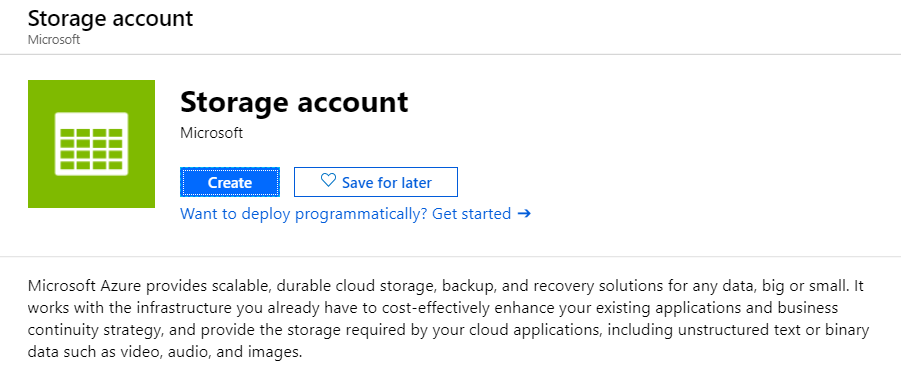
* Have access to an Azure subscription (even a [free trial](https://azure.microsoft.com/en-gb/free/))
* Have sufficient access on this subscription to create resources
* Have access to a Service Principal (or the access to create one)

The above aren’t show-stoppers and can be worked around – but we haven’t taken them into account in the instructions below!

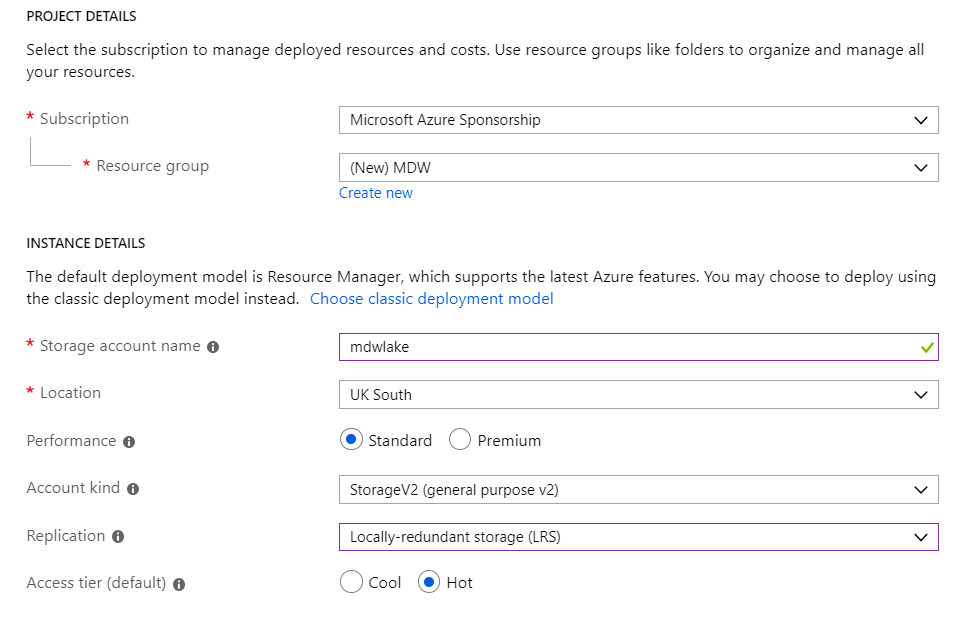
## Lab 01.A – Creating A Data Lake

First we need to create a new storage account to house our lake. Unlike the first generation of Microsoft Data Lakes, this is not a separate resource and is created through the normal blob process.

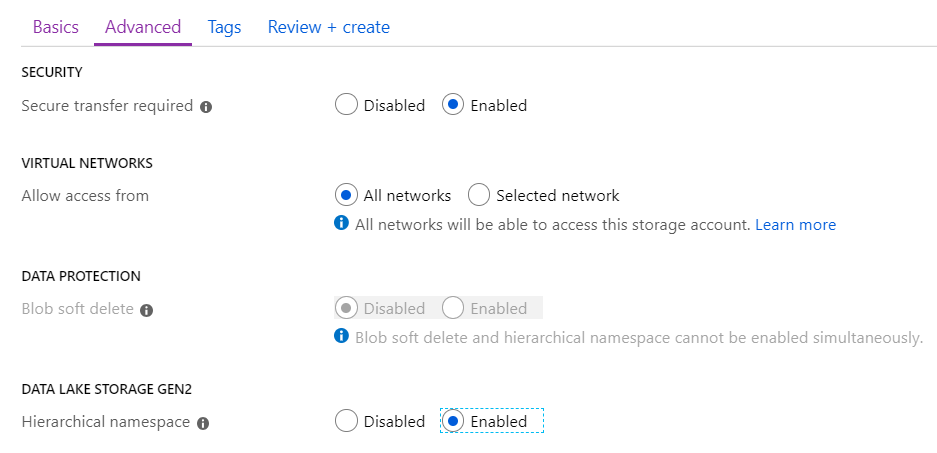
1. In the Portal, click “Create a Resource” and search for “Storage”. You should find the following.



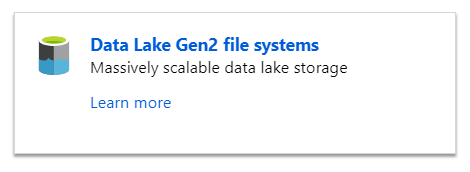
1. Click “Create” and you’ll see a short form to fill in. Bear the following in mind:
   1. I’ve created a new resource group for the labs – but you can use any!
   2. The blob name has to be unique – not just for you, but across all blobs
   3. Pick your region carefully – you want all the items you create to live in the same region otherwise you’ll incur egress costs
   4. Make sure you choose “StorageV2” – this is the only version that Data Lake functionality is enabled for
   5. I’ve kept things cheap by turning off geo-replication – I don’t need backup copies of my lake, I would definitely turn this on for a production environment!



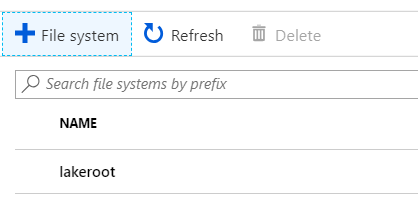
1. You’ll then want to click the “Advanced” button, where we will see the option to enable “Hierarchal Namespace” – this is essentially making it a Data Lake Store Gen 2 resource



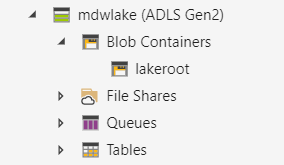
1. You can now click “Review and Create” and finally “Create” once it has validated your request. This will now create a blob storage account for you with that all important hierarchal namespace enabled. It might take a few minutes.
2. Once complete, you can navigate to the resource and you should see the all important ADLS Gen 2 box:



1. Click into this box and you’ll be prompted to click “+ File System” to create a new lake structure. Do this now.
2. I’ve called mine “lakeroot” so I can now see the following:



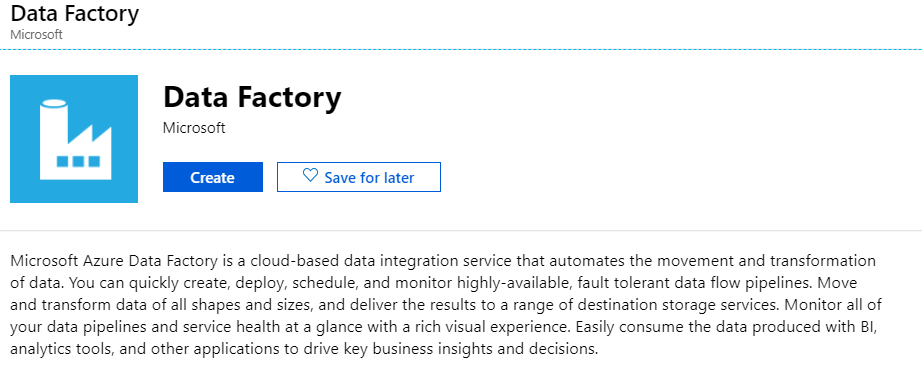
1. If you click into your new file structure, you’ll see a large Data Lake icon and a message telling you to use “Azure Storage Explorer” to interact with the lake from now on. Make sure you have this installed locally.
2. If you have not used Storage Explorer before, you’ll need to enter your Azure Subscription credentials. Once you do, you should see a list of your storage accounts and the new hierarchal namespace will appear under blob containers. Mine looks as follows:



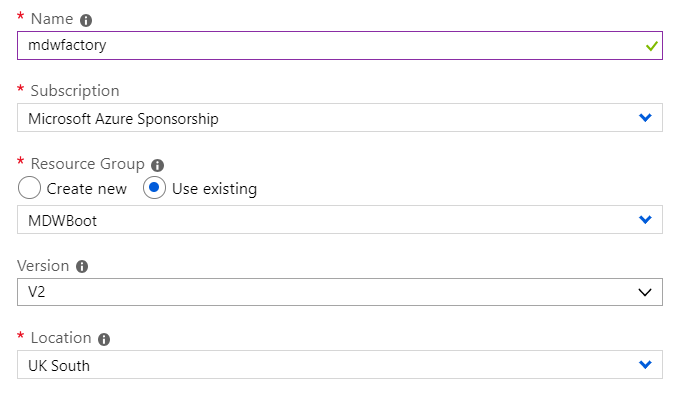
## Lab 01.B – Creating A Data Factory

Now that we have a lake ready to use, we can setup Data Factory to start copying data into it.

1. Use the “Create a Resource” button again and this time search for Data Factory to find the following:

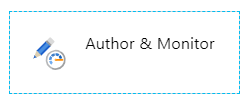


1. Click create and follow the instructions



For simplicity, I’ve not enabled git – but you can do if you have the details ready to go.

1. Navigate your new data factory, and click on the “Author and Manage” button to open the Data Factory workspace



1. You’ll see the entry screen load, but we want to go straight to our pipeline creation, so click on the  button, or the “Create Pipeline” button:

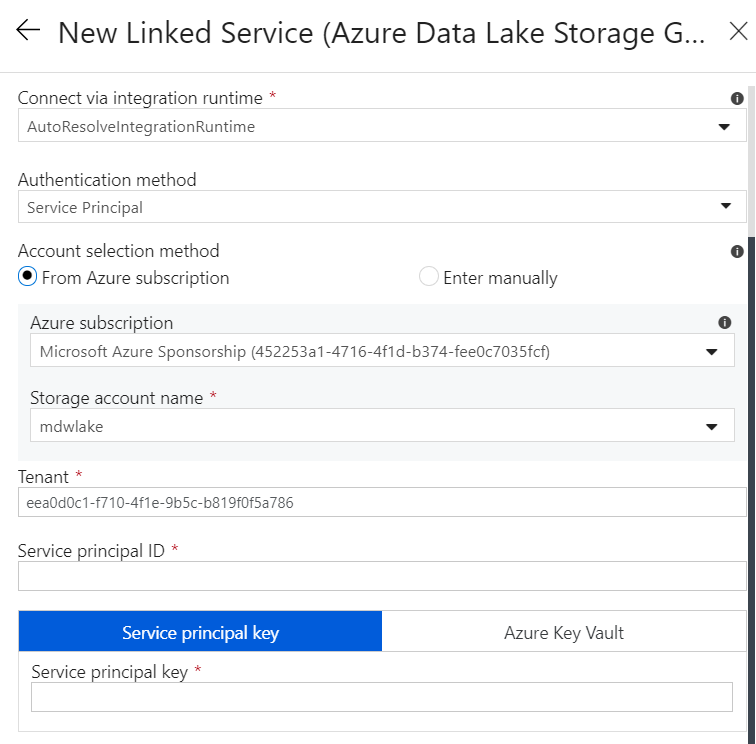


1. Our first step is to ensure the Data Factory has access to the lake, so we can go straight to the  button.
2. Click on the +New button and select the Data Lake Store Gen 2 icon:



This will open up a linked services blade, specifically for ADLS Gen 2.

1. We could use a blob key, but it’s better if we allow use a service principal account to do this. We could allow the ADF Managed Service Identity to have access – this is a special service account managed by data factory itself, but I prefer to have different service principals to enable more fine grain on access levels.
2. Select Service Principal and you’ll be prompted for further details:

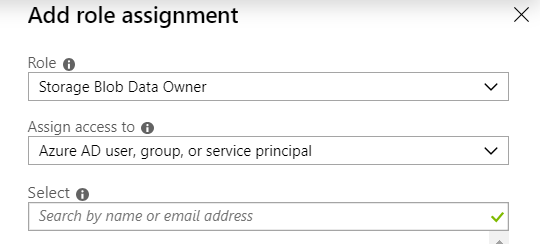


Now, before we go any further, we haven’t actually given our Service Principal access to our lake yet! Before you create your linked service in data factory, let’s make sure it has access. Don’t close your data factory window as we’ll be coming back here later!

If you haven’t got a Service Principal yet – follow the optional lab 1.1!

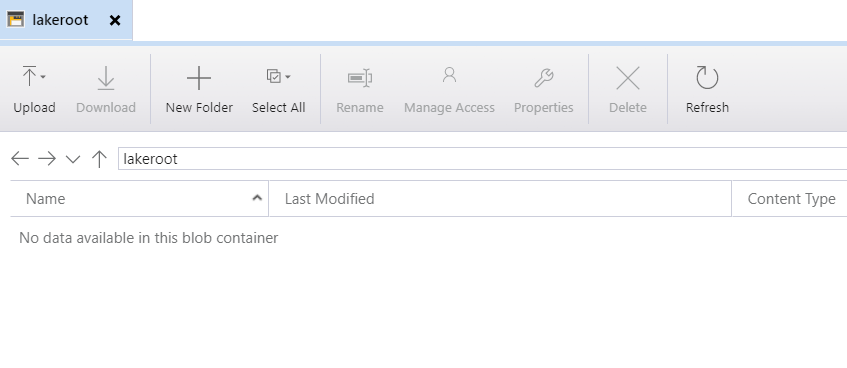
1. Back in the Azure Portal, navigate to the Storage Account you set up. Click on the  button.
2. We want to create a new role assignment for our lake user, so click on “Add” or the “Add a Role Assignment” button.

We want to create a new “Storage Blob Data Owner”, which gives our Service Principal access to read and write data in the lake.



*Note: You should be able to use the more restrictive roles (storage blob data reader/writer) but ADF sometimes errors with these roles – we assume this will be fixed over time!*

1. Switch back to Azure Storage Explorer and navigate to your lake namespace



You’ll see many of our options are greyed out, this is because we need to start off with some objects.

1. We can then use the +New Folder button to create the base set of data lake folders:

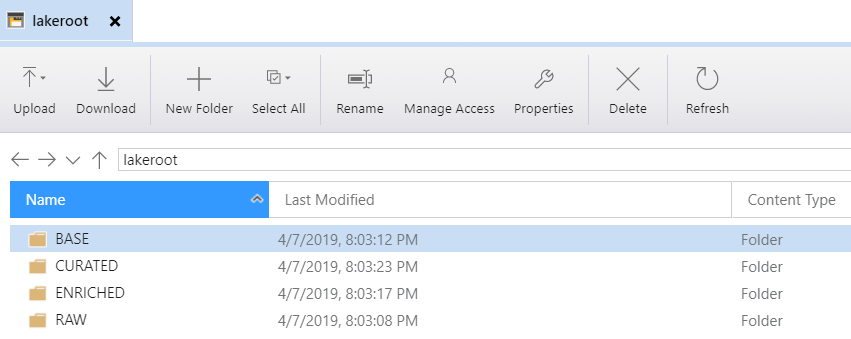
RAW

BASE

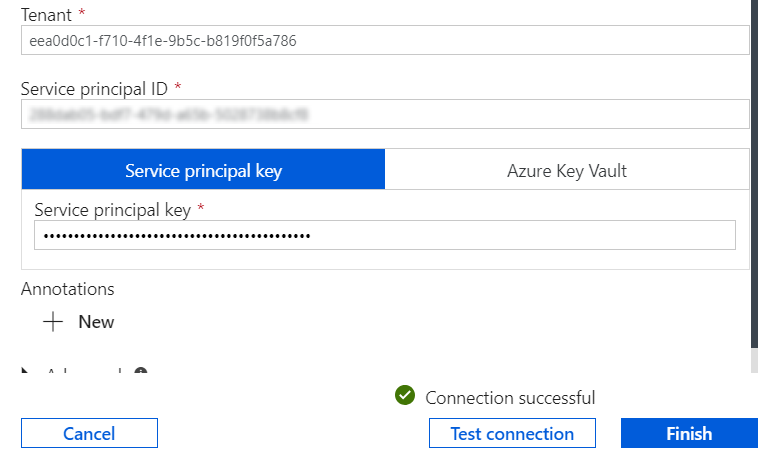
ENRICHED

CURATED

1. Your lake should now look as follows:



1. We can now return to Data Factory to finish setting up our Linked Service.
2. Enter your Service Principal ID and Key then click the “Test Connection” button to ensure access has been set up correctly:



1. If you see the “Connection Successful” button, you can click “Finish” and you now have a Data Factory set up to work with your Data Lake.

