Lab 1 – Create Azure resources

Welcome to Lab 1!

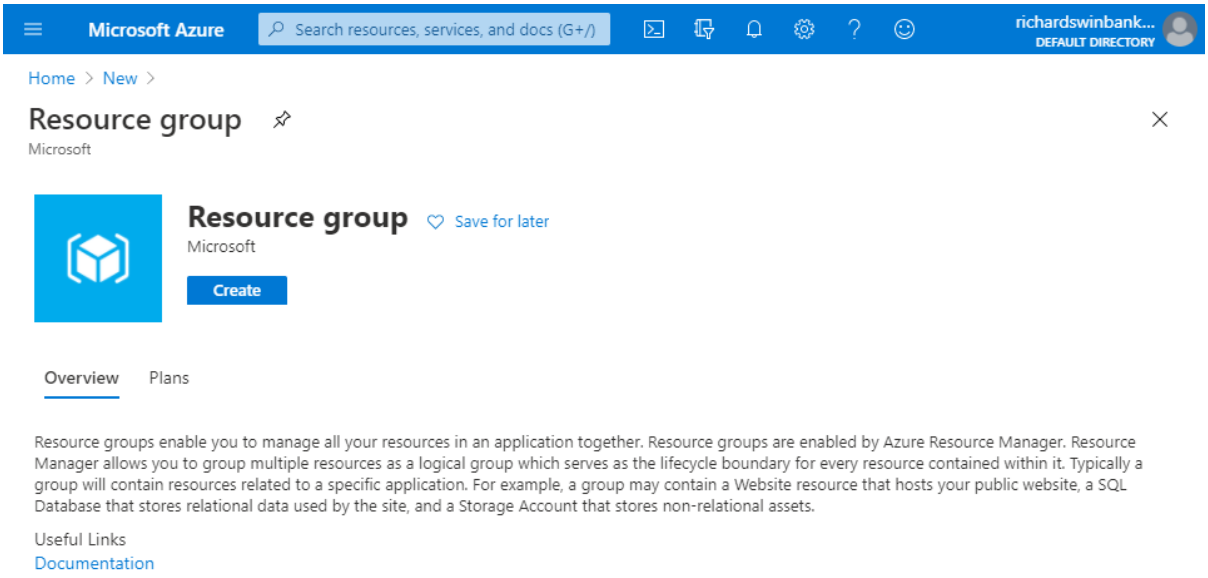
To complete the labs, you're going to need an Azure subscription (if you don't have one, you can sign up for a free trial). Whichever subscription you choose, you'll need enough access to create resources in it.

Lab 1.1 – Create a resource group

Resource groups are logical containers for resources in Azure. In this lab you will create a resource group to contain all the resources you create in later labs. This will make cleaning up easy – when you've finished all the labs, you can just delete the resource group.

1. In the Azure Portal, click "Create a resource" and search for "Resource group".

2. On the "Resource group" overview, click "Create".



3. Give the resource group a name, and choose the Region geographically closest to you.

4. Click "Review + create", then "Create".

Lab 1.2 – Create data lake storage

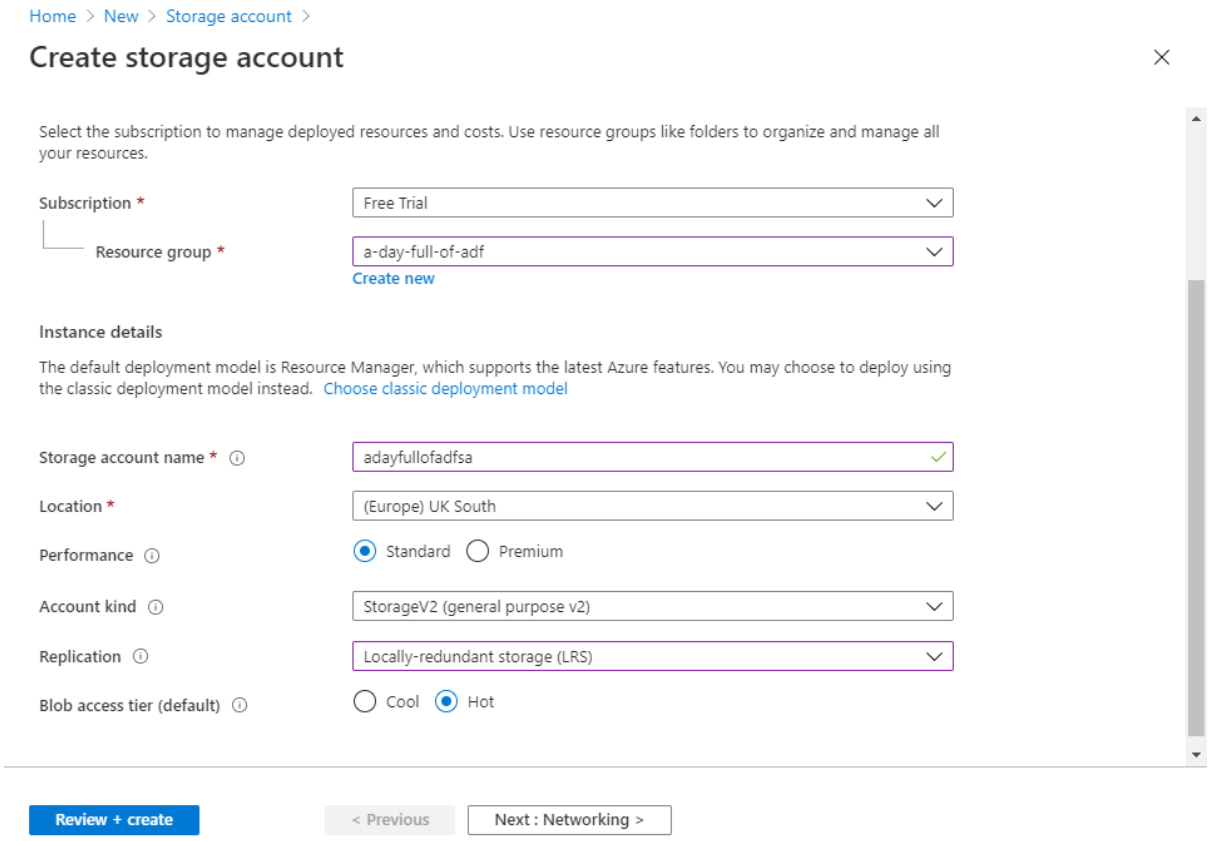
Data lake storage is blob storage in an Azure storage account, with an important feature: hierarchical namespaces are **enabled**. This makes certain file operations – renaming file folders, for example – much more efficient.

1. In the portal, click "Create a resource" and search for "Storage account". Click "Create" on the overview screen.

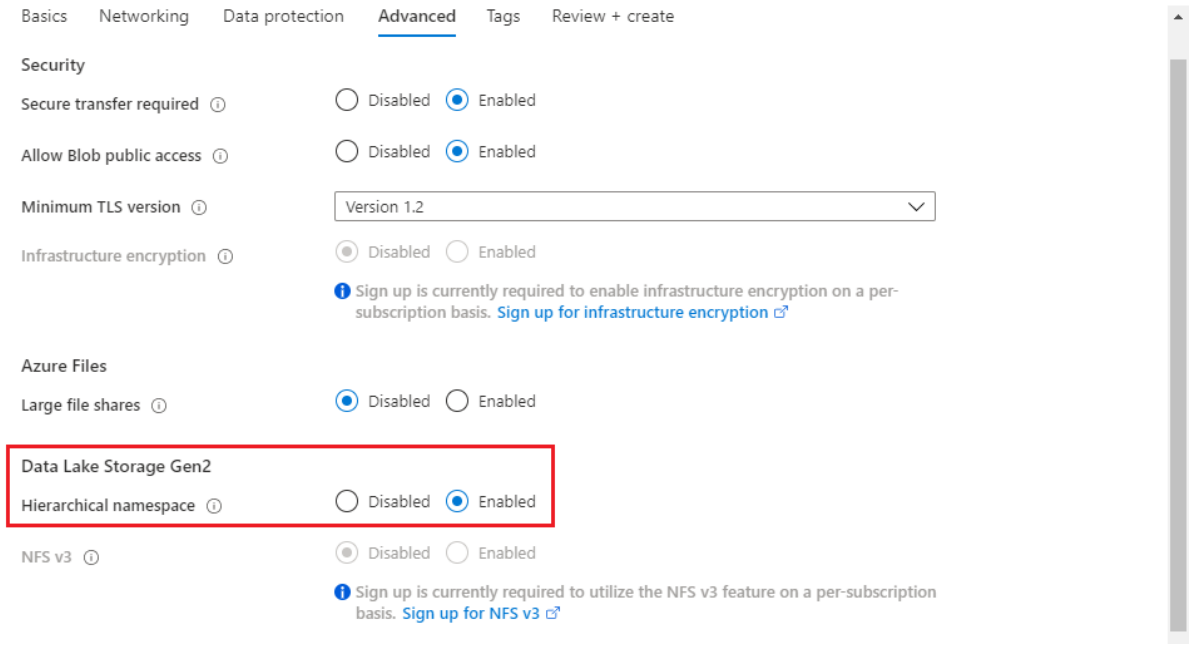
2. Complete the **Basics** tab like this:

   - Choose your subscription and the resource group you created in Lab 1.1.
   - Enter a storage account name – this must be globally unique (across the entire Azure platform).

- Choose the same location you specified for your resource group. Having storage located close to you makes for faster data transfers.
- Choose account kind “StorageV2” – this is the only kind that supports hierarchical namespaces.
- Choose replication option “Locally-redundant storage”. This is nice and cheap for lab work, but you’ll want something more resilient in a production environment!



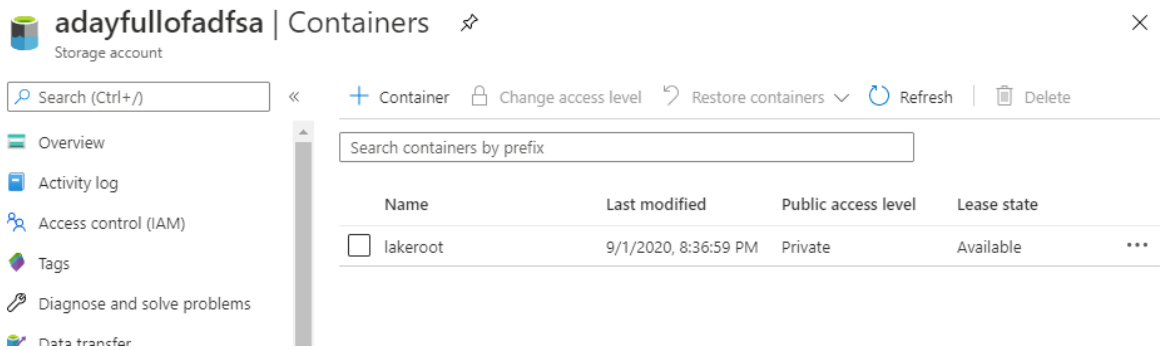
3. On the **Advanced** tab (click through Networking and Data protection to get there), enable hierarchical namespaces. This step is **essential** to make this storage account a data lake.



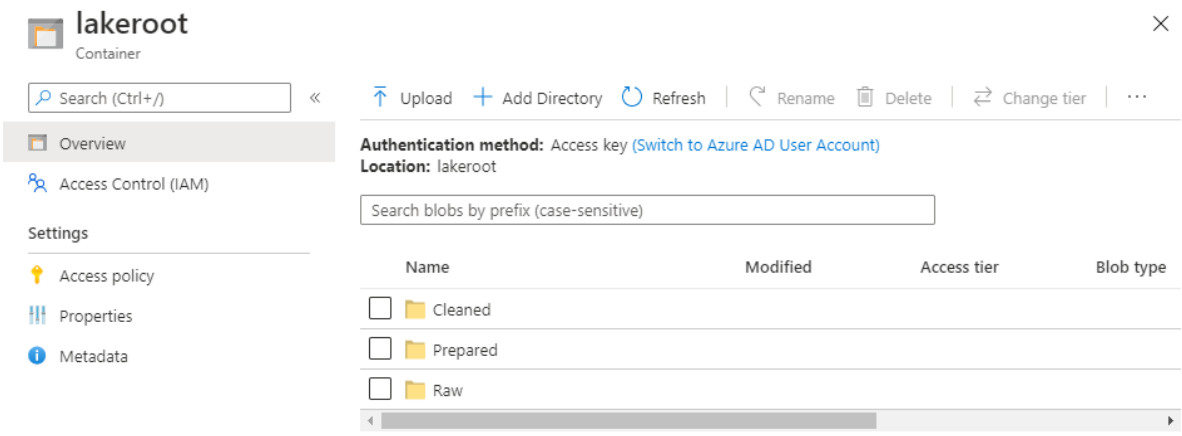
4. Click “Review + create”, then “Create”. When the data lake finishes deploying (this may take a couple of minutes), click on “Go to resource”.

   The storage account blade displays tiles for each of the four supported storage types. If your “Containers” tile looks like the one on the left, you’re in business. (If it looks like the one on the right, you forgot to enable hierarchical namespaces – delete the storage account and have another go).

5. Click on the data lake containers tile to open the (empty) list of containers. Use the “+ Container” button to create a container with the name “lakeroot”. After creation, the container appears in the list.



6. Click on the new “lakeroot” entry to open the container. The menu bar above the list now contains a “+ Add Directory” button – use this to create three directories in the container: “Raw”, “Cleaned” and “Prepared”.



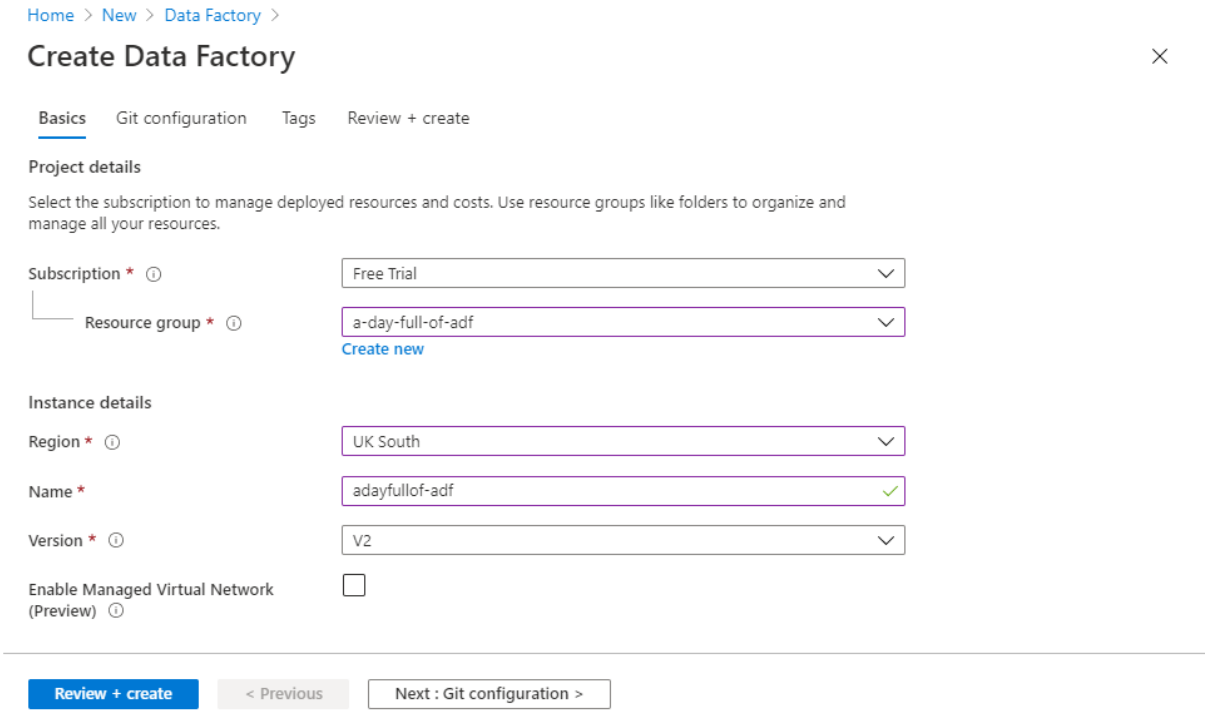
Lab 1.3 – Create an Azure Data Factory

The main event! It’s time to create your Azure Data Factory instance.

1. In the portal, click “Create a resource” and search for “Data Factory”. Click “Create” on the overview screen.

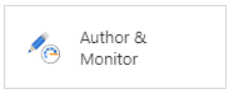
2. Complete the **Basics** tab like this:

   - Choose your subscription and the resource group you created in Lab 1.1.
   - Choose the same location (region) you specified for your storage account. **This has cost implications** – transferring data from a storage account in one region to a data factory in another incurs an outbound data transfer charge.
   - Enter a data factory name – this must be globally unique.
   - Choose version V2.



3. If you have a Git repository available, you can connect your new factory to it on the **Git configuration** tab. Don't worry if you don't have a repo ready – just click "Configure Git later".

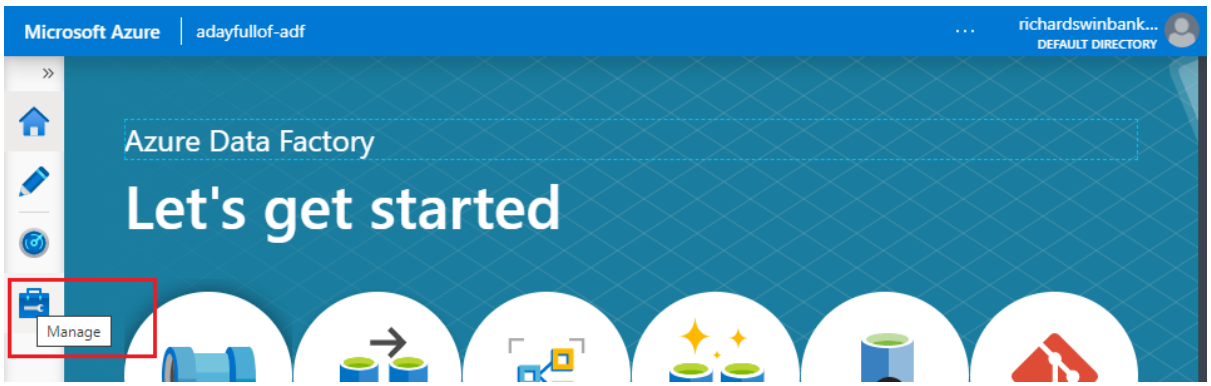
4. Click "Review + create", then "Create". When the factory deployment is complete, click on "Go to resource", then on the "Author & Monitor" tile to launch the ADF User Experience (ADF UX).



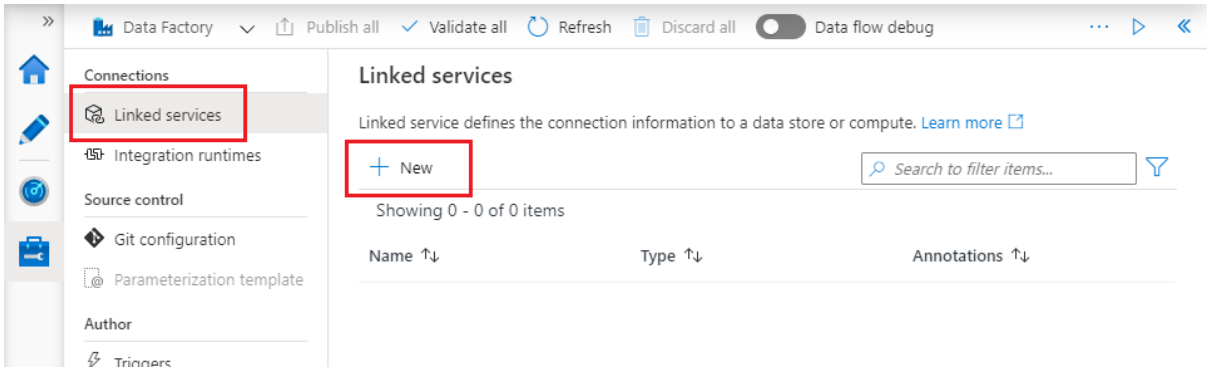
Lab 1.4 – Connect to the data lake

To enable ADF pipelines to use data in the lake you will need a data factory **linked service** connection.

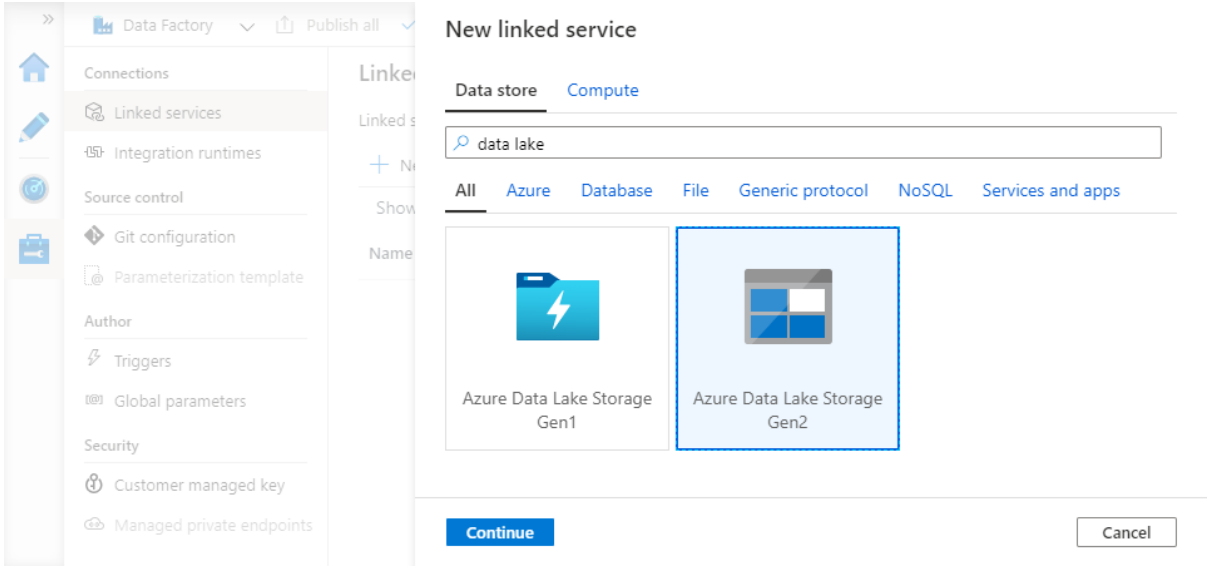
1. In the ADF UX's leftmost sidebar, click the "Manage" button to open the Management Hub.



2. Select the "Linked services" item from the "Connections" section of the Management Hub sidebar, then in the main Linked services pane click "+ New".



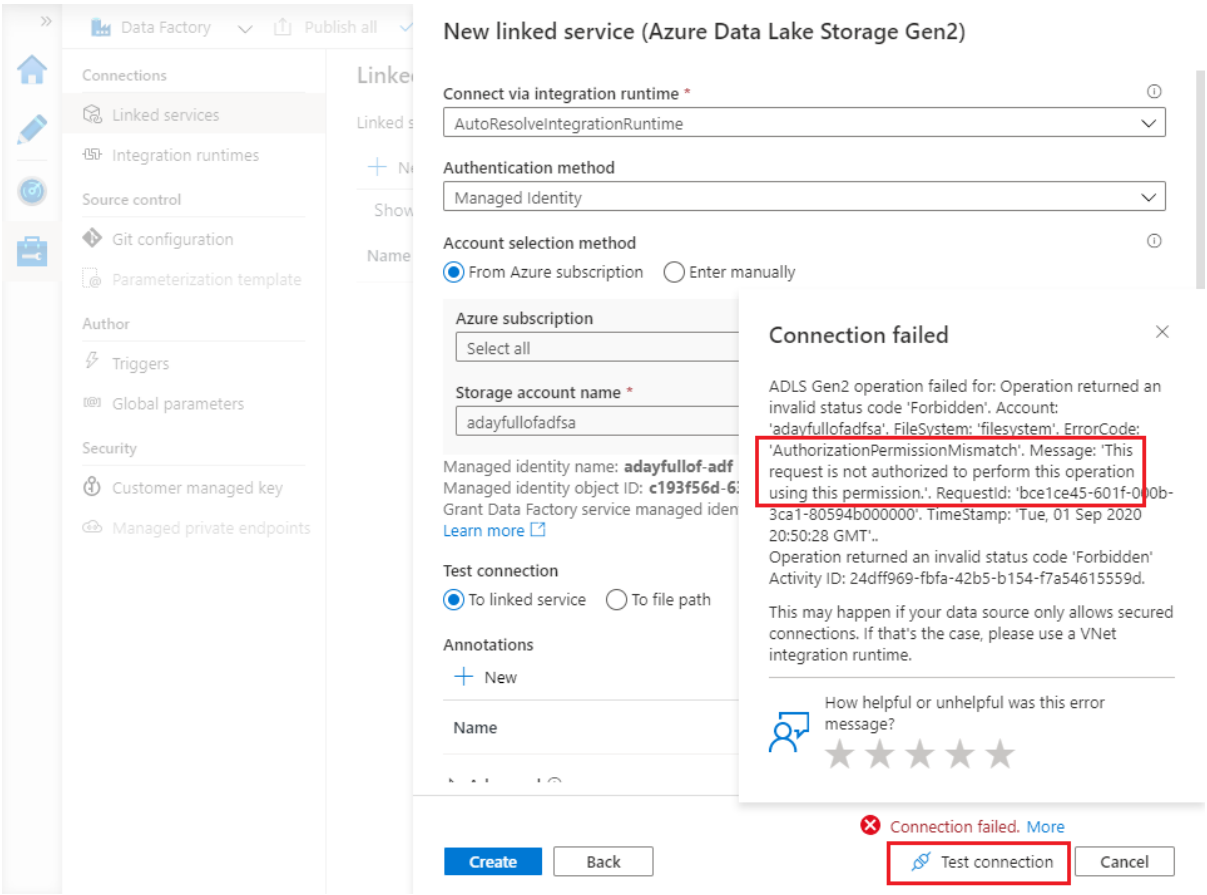
3. Search for "data lake", then choose "Azure Data Lake Storage Gen2" and click "Continue".



4. Configure linked service details on the "New linked service (Azure Data Lake Storage Gen2)" blade like this:

- Give it a name.
- Choose Authentication method "Managed Identity". The default method ("Account key") requires extra work to pass keys around securely – "Managed Identity" uses an Azure Active Directory service principal created automatically for your data factory when you created it. This service principal is called a Managed Service Identity (MSI).
- Choose your data lake storage account from the "Storage account name" dropdown.

5. Click “Test connection” at the bottom of the blade. The connection test will fail, because the factory’s MSI does not have access to the data lake yet – you will receive an error message like “This request is not authorized to perform this operation using this permission”.



6. You will grant the necessary access in a moment – for now, just click “Create” to create the linked service.

Lab 1.5 – Grant access to the data lake

You can manage access to Azure resources in the Azure portal. Open a new browser tab to allow you to keep the ADF UX open.

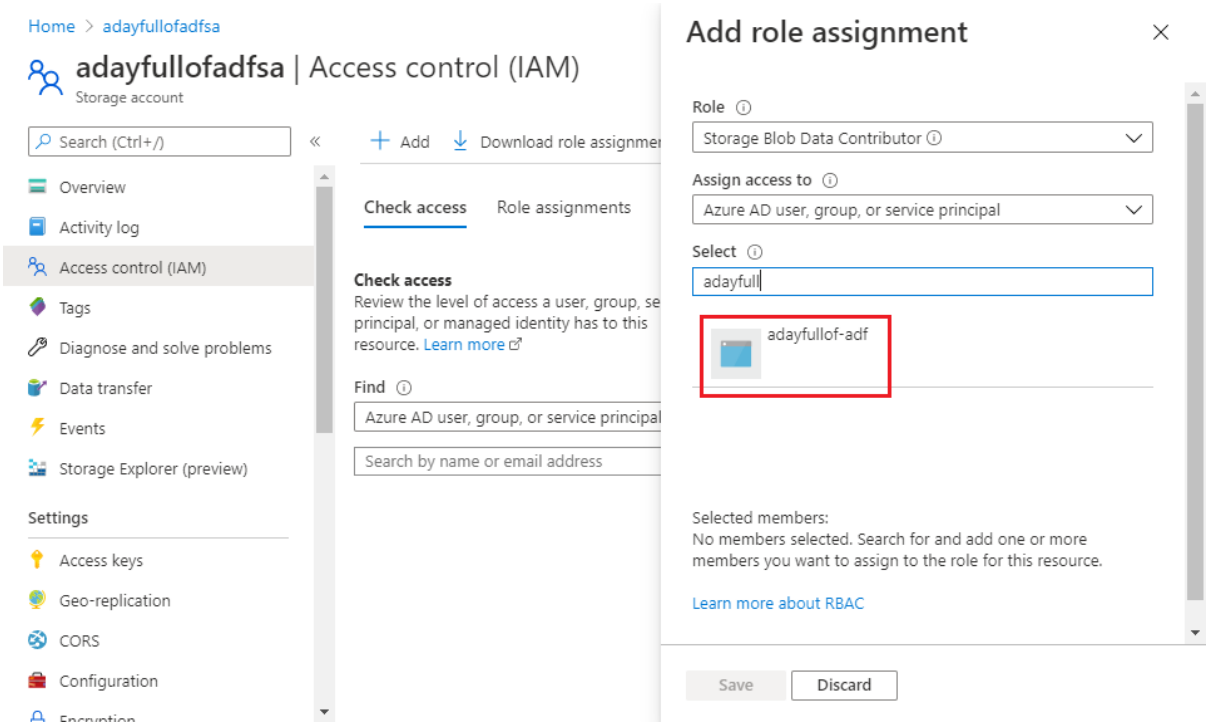
1. Browse to your data lake resource blade – you can find it in the list of resources on the portal home page, or by selecting your resource group to see resources inside it, or by using the search box in the portal’s top menu bar.

2. Click “Access control (IAM)” in the storage account (data lake) resource blade.

3. In the “Add a role assignment” tile, click the “Add” button.

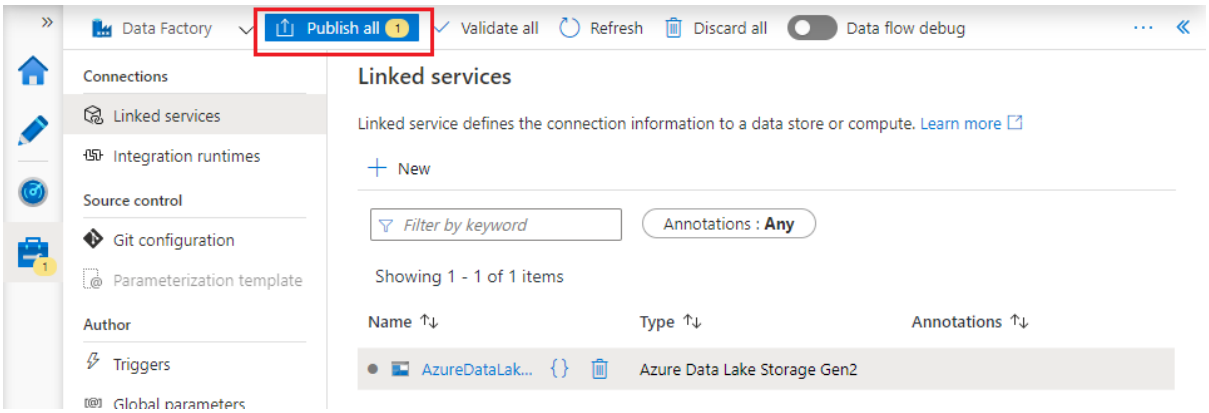
4. In the “Add role assignment” blade, choose the “Storage Blob Data Contributor” role – this authorises read, write and delete access in your data lake.

5. Enter the name of your data factory in the “Select” text box to search for the MSI. (The MSI’s display name is the same as the data factory name). When the MSI appears below the “Select” box, click to select it, then click “Save”.



7. Return to the ADF UX Management Hub, and click on your data lake linked service to re-open the editing blade. Click “Test connection” again, and verify that this time the connection test succeeds. If the test fails, wait a few minutes, then try again – it may take a short time for permission changes to take effect). Click “Cancel” to close the editing blade.

8. If you have not enabled Git for your data factory, the only way to save your changes is to publish them – click “Publish all” in the top menu bar to do so. (If Git is enabled, linked service changes are saved automatically to your Git repo).



Recap

In Lab 1 you:

- created an Azure resource group
- created data lake storage
- created an instance Azure Data Factory
- created and authorised a connection from the factory to your data lake storage.