Lab 1 – Create Azure resources

Welcome to Lab 1!

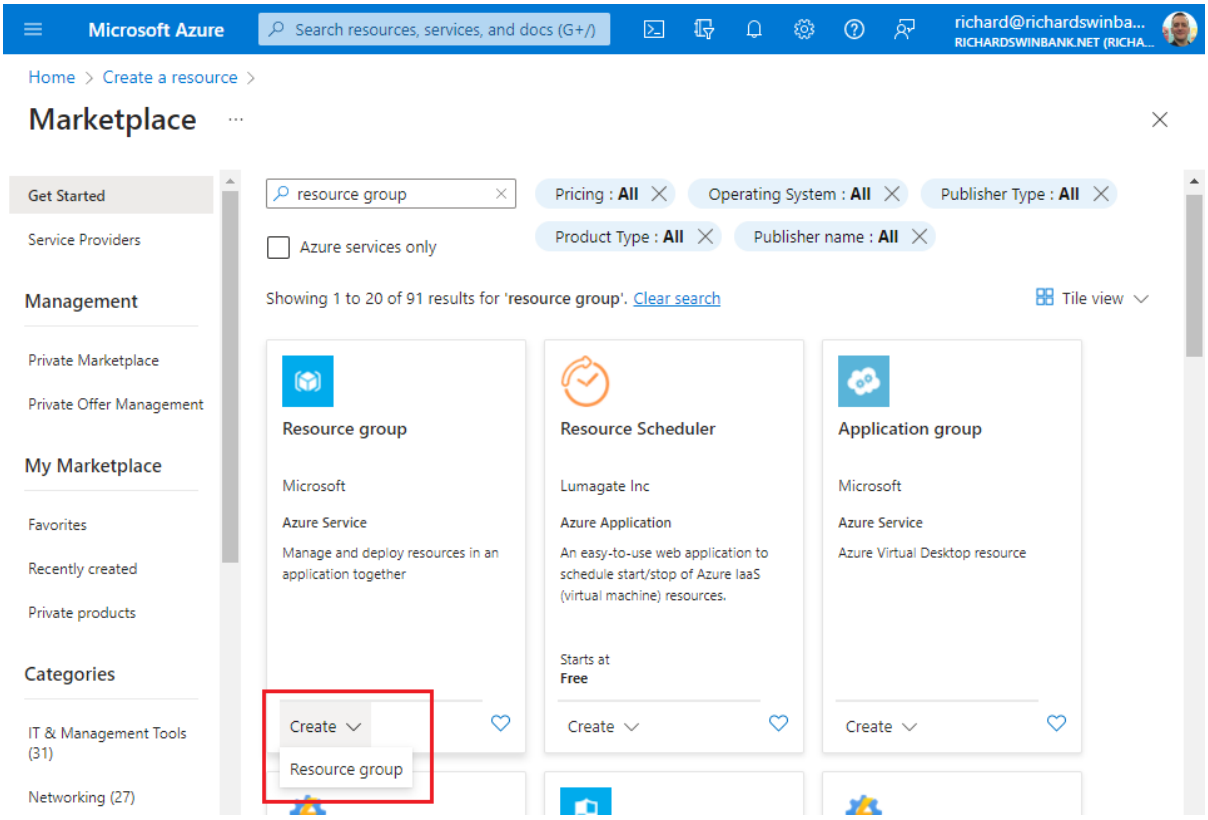
To complete the labs, you're going to need an Azure subscription (if you don't have one, you can sign up for a free trial). Whichever subscription you choose, you'll need enough access to create resources in it.

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 build in later labs. This will make cleaning up easy – when you've finished all the labs, you can just delete the resource group.

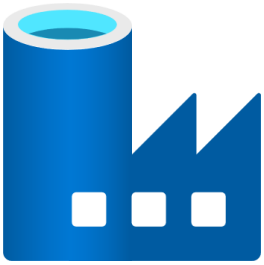
1. In the Azure Portal, click "Create a resource" and search for "resource group".

2. In the "Marketplace" search results, find the "Resource group" tile, then use its "Create" dropdown to choose the "Resource group" option:



3. On the "Create a resource group" blade, choose your subscription and give the resource group a name. Choose the Region geographically closest to you.

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



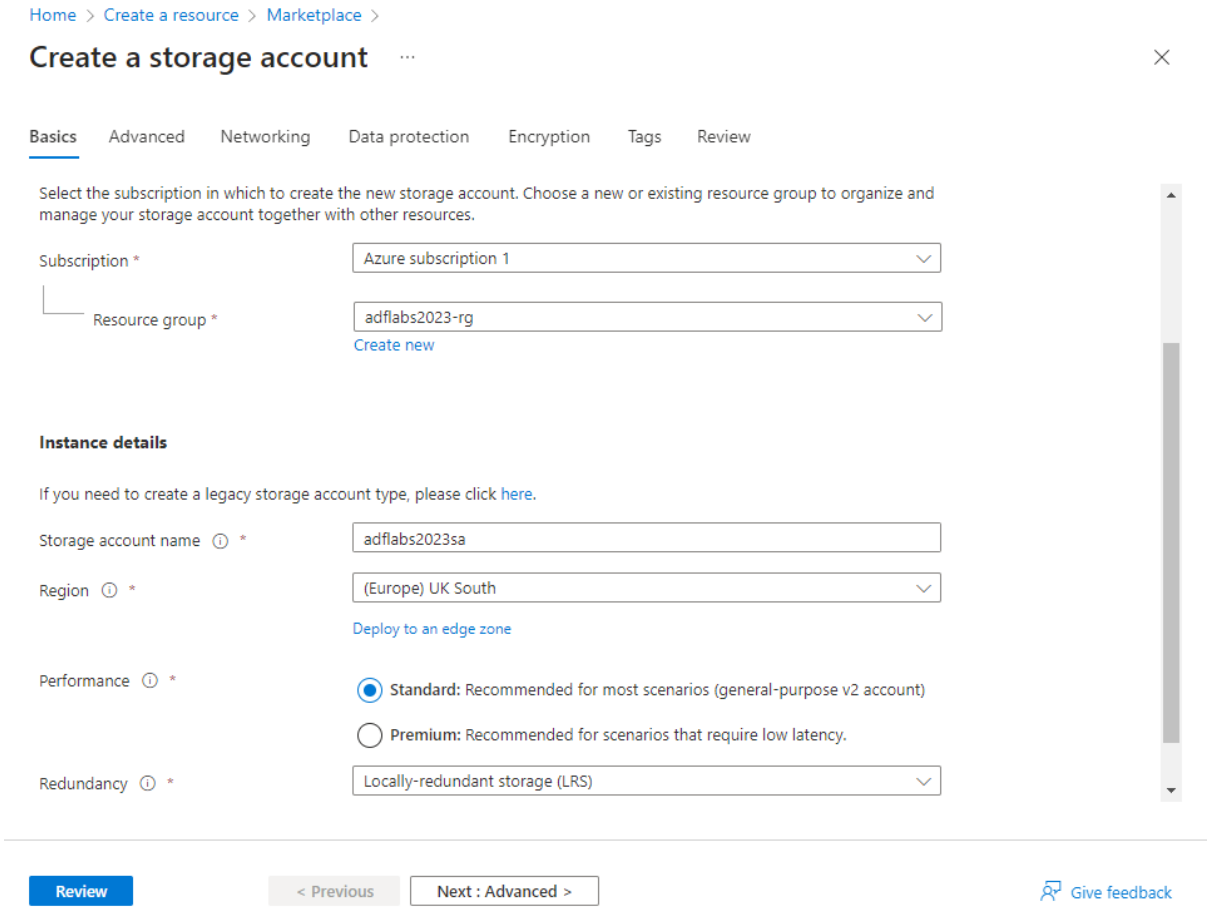
Lab 1.2 – Create data lake storage

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

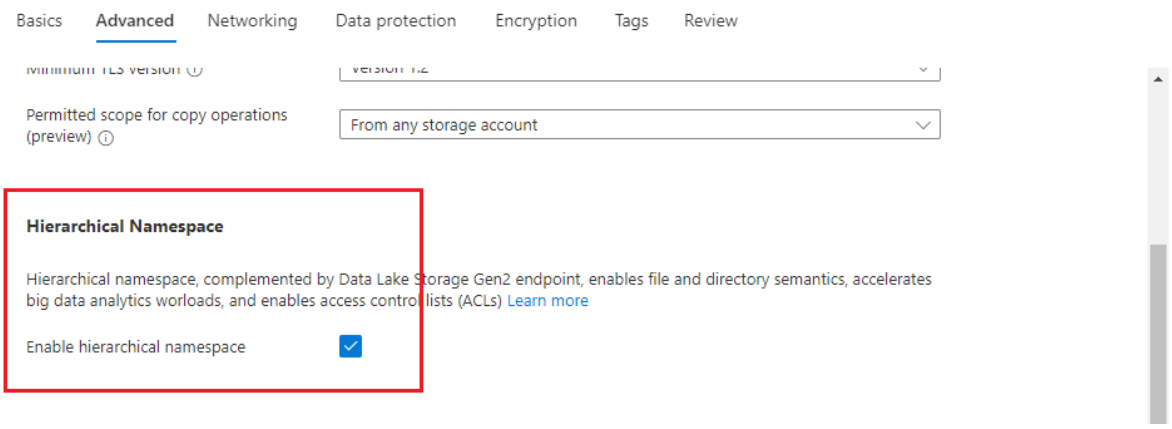
1. In the portal, click “Create a resource” and search for “storage account”. Find the “Storage account” tile in the “Marketplace” search results, then use its “Create” dropdown to start configuring a new storage account.

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 cloud).
   - Choose the same **region** you specified for your resource group. Having storage located close to you makes for faster data transfers.
   - Choose “Standard” **performance**.
   - Choose **redundancy** option “Locally-redundant storage (LRS)”. (This is nice and cheap for lab work, but you’ll want something more resilient in a production environment!).

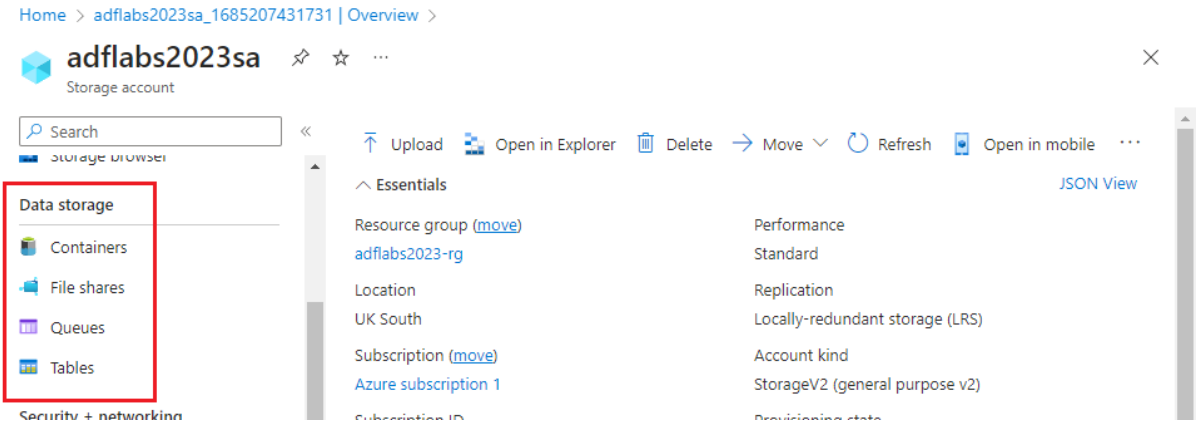


3. On the **Advanced** tab, enable hierarchical namespaces. This step is **essential** to make this storage account a data lake.



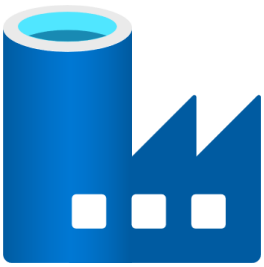
4. Click “Review”, then “Create” – you may need to wait a few moments for automatic validation to complete before the “Create” button becomes available. When the data lake finishes deploying (this may take a couple of minutes) click “Go to resource”.

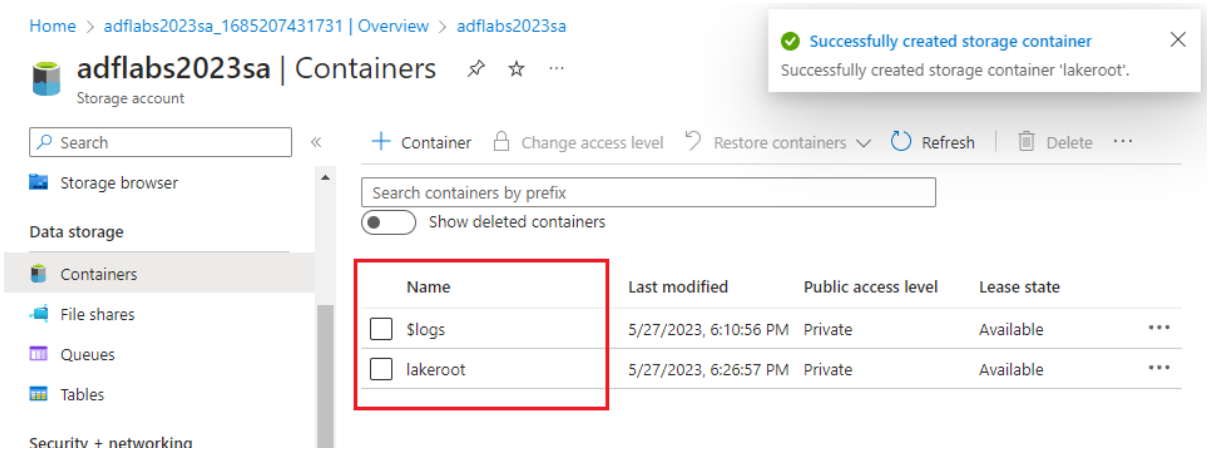
   The “Data storage” section of the storage account blade sidebar menu contains an option for each of the four supported storage types:



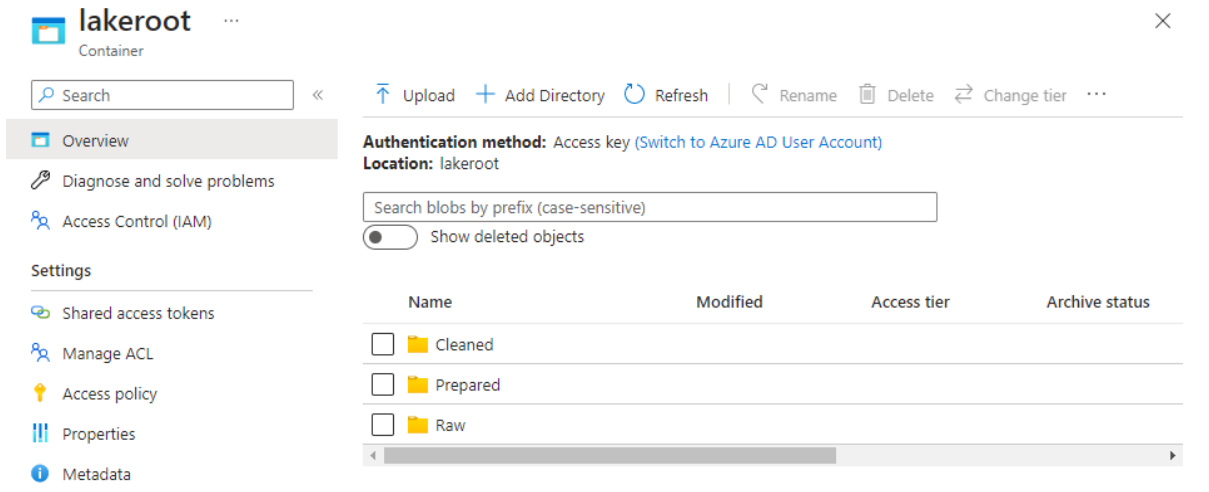
   If the icon for the “Containers” button looks like the one on the left (below), you’re in business. If it looks like the one on the right, you forgot to enable hierarchical namespaces – delete the storage account and have another go.

5. Click on the data lake “Containers” menu option to open the list of containers. A single container called “$logs” is already present, created automatically to contain log files for the storage account itself. Use the “+ Container” button to create a second, private 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".



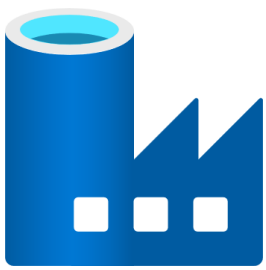
Lab 1.3 – Create an Azure Data Factory

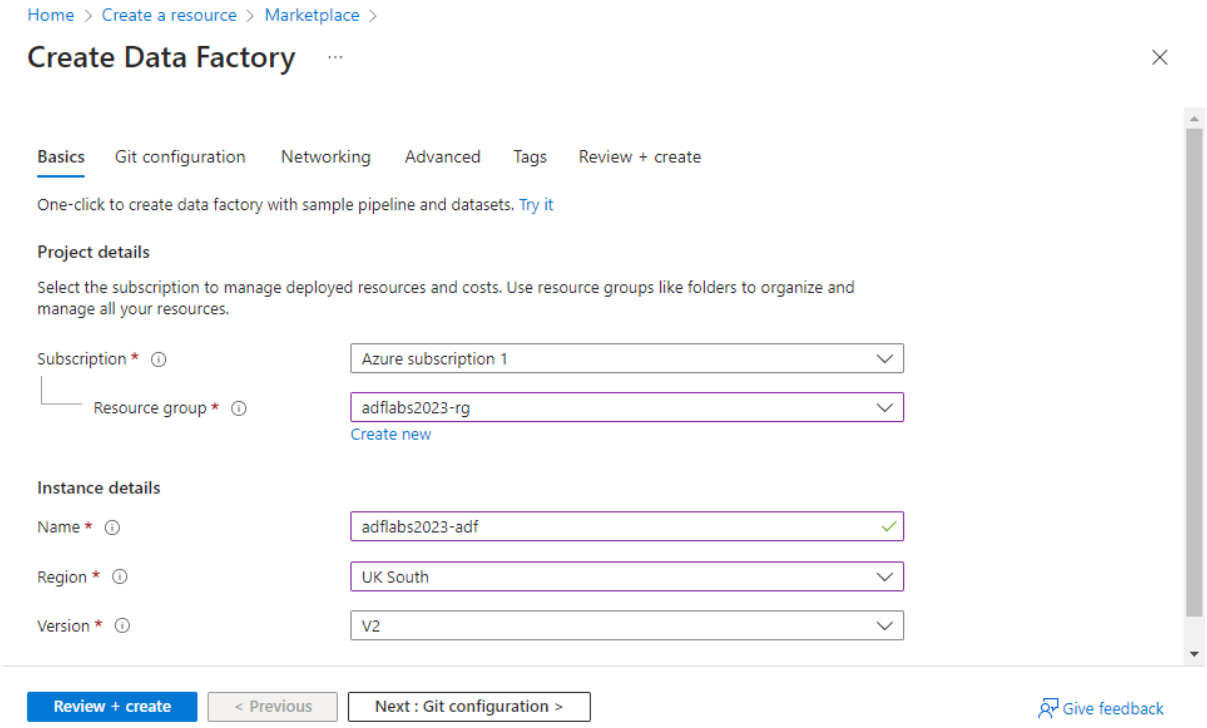
The main event! It's time to create your Azure Data Factory instance.

1. In the portal, click "Create a resource" and search for "data factory". Find the "Data Factory" tile in the "Marketplace" search results, then use its "Create" dropdown to start configuring the new factory instance.

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

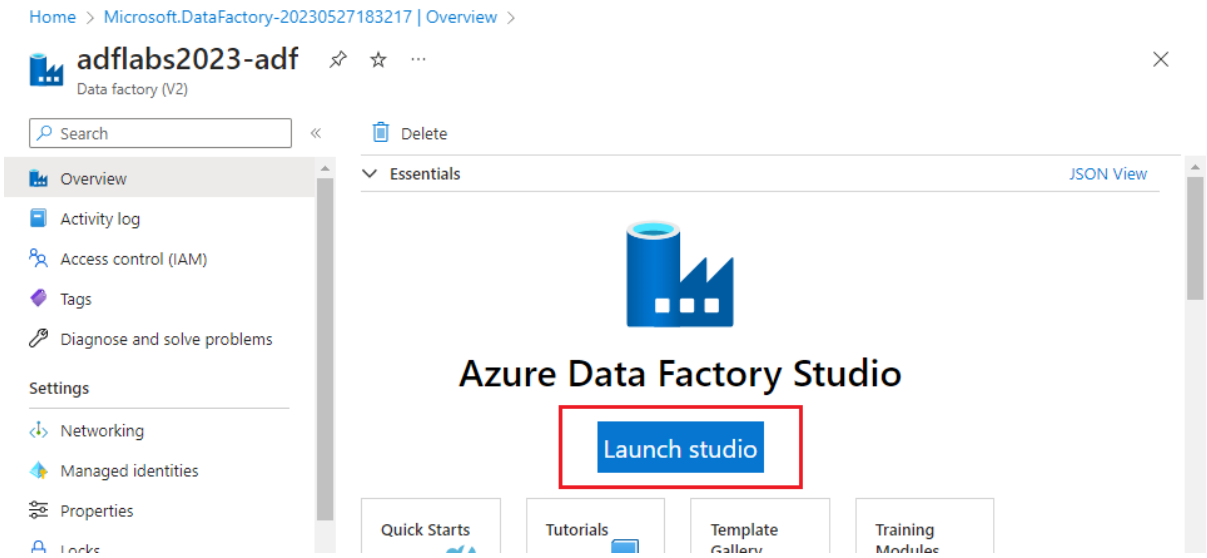
   - Choose your **subscription** and the **resource group** you created in Lab 1.1.
   - Enter a data factory **name** – this must be globally unique.
   - Choose the same **region** you specified for your storage account. **This has cost implications** – transferring data from a storage account in one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





3. You can connect your new factory to a Git repository on the **Git configuration** tab, but for the purposes of these labs we will omit that step – make sure that the "Configure Git later" checkbox is ticked.

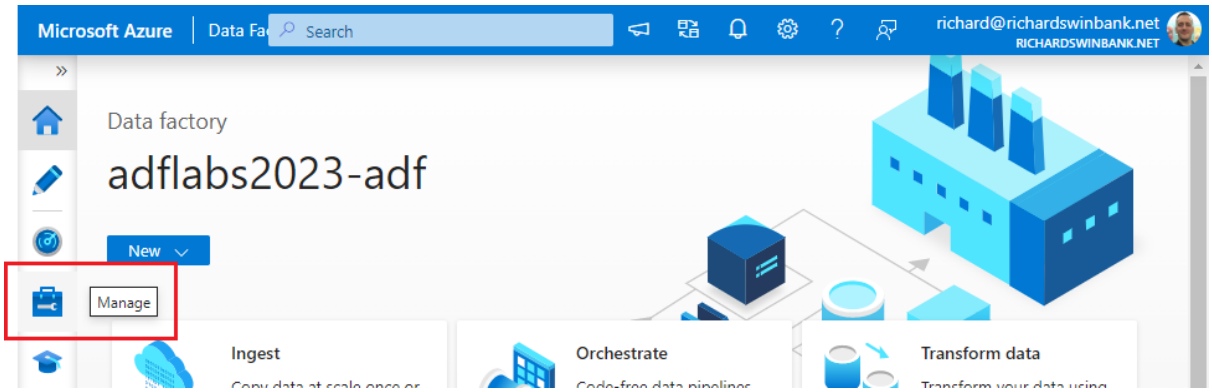
4. Click "Review + create", then "Create". When the factory deployment is complete, click on "Go to resource". On the factory's portal Overview page, launch Azure Data Factory Studio to connect to your new factory.



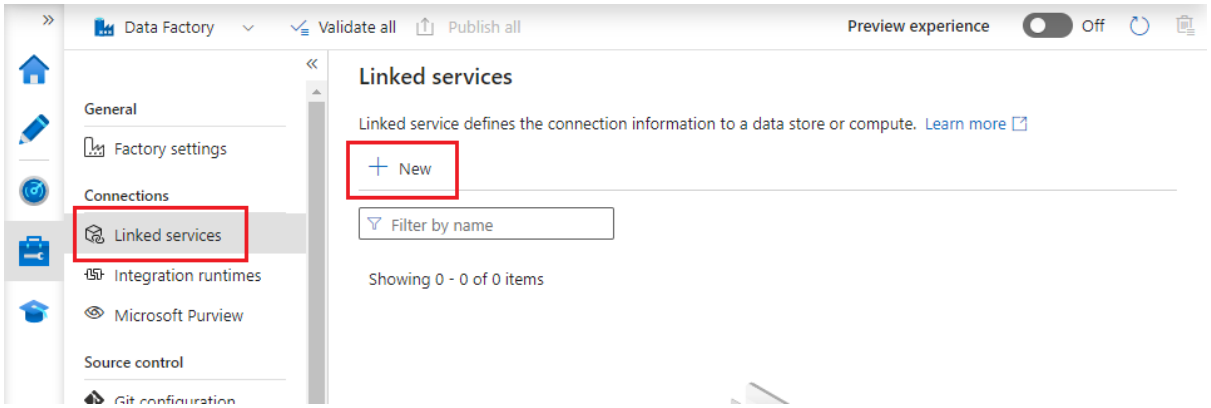
Lab 1.4 – Connect to the data lake

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

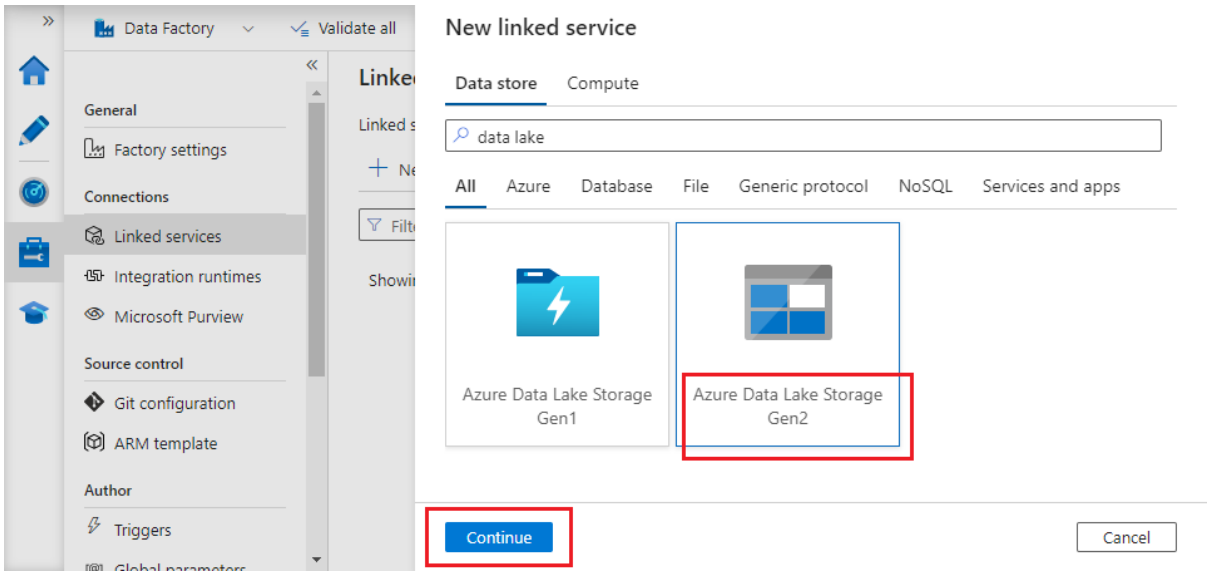
1. In the ADF Studio sidebar, click the "Manage" button to open the Management Hub.



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



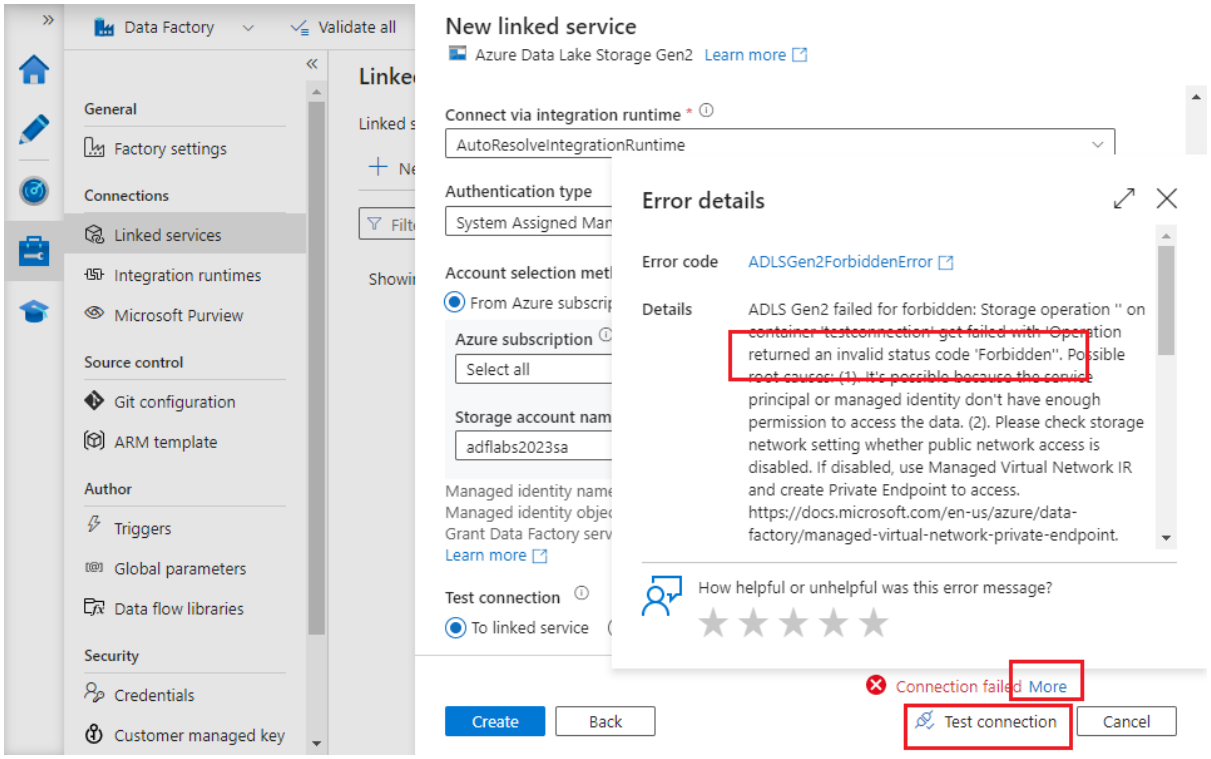
3. Search for "data lake", then choose "Azure Data Lake Storage Gen2" and click "Continue".



4. Configure linked service details on the "New linked service" flyout like this:

- Give it a **name**.
- Choose **Authentication type** "System Assigned Managed Identity" – this uses an Azure Active Directory service principal, created automatically for your data factory when you deployed it. This type of service principal is called a Managed Identity.
- 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 – this is to be expected! Use the "More" link next to the "Connection failed" message – the error details will indicate that the test operation "returned an invalid status code 'Forbidden'". This is because the factory's managed identity does not have access to the data lake yet.



6. You will grant the necessary access in a moment – for now, just close the Error details popup and click "Create" to create the linked service.

Lab 1.5 – Grant access to the data lake

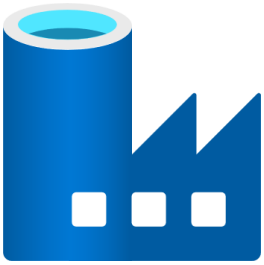
You can manage access to Azure resources in the Azure portal. Open a new browser tab so that you can keep ADF Studio open – unsaved changes are discarded when you quit ADF Studio.

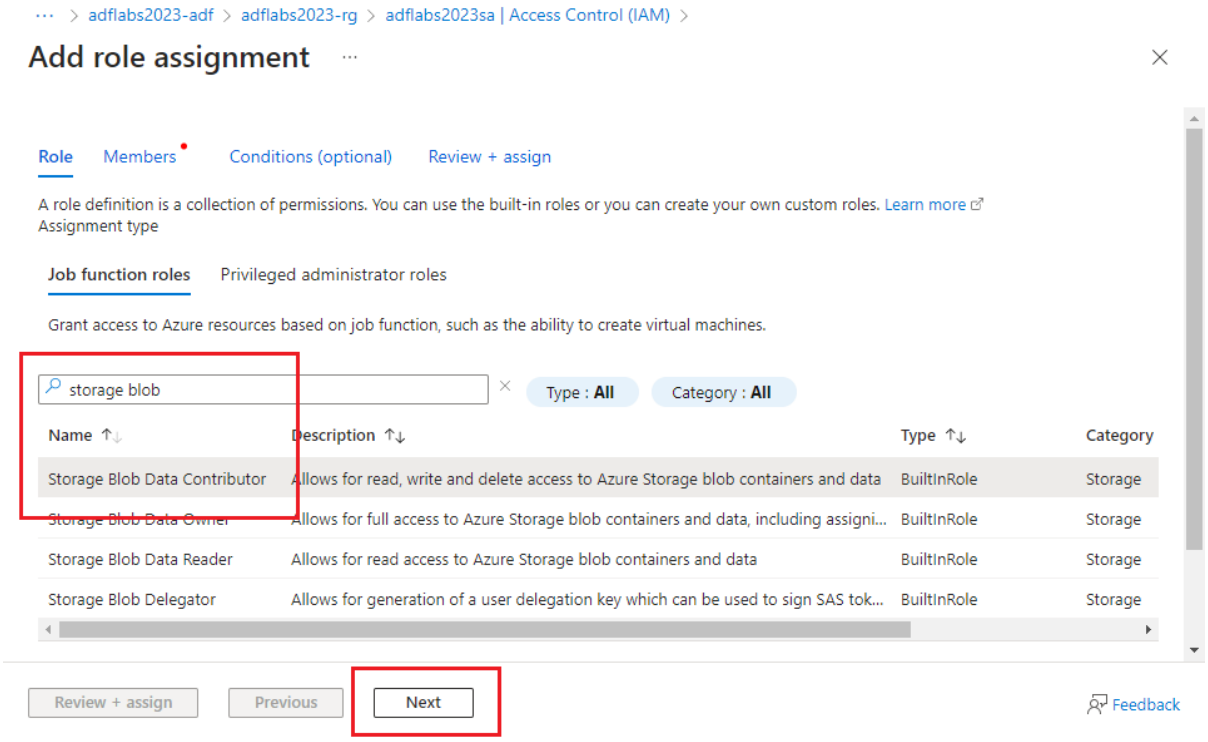
1. Browse to your data lake resource blade – you can find the lake in the list of recent resources on the portal home page, or by selecting your resource group to see the resources it contains, or by using the search box in the portal's top menu bar.

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

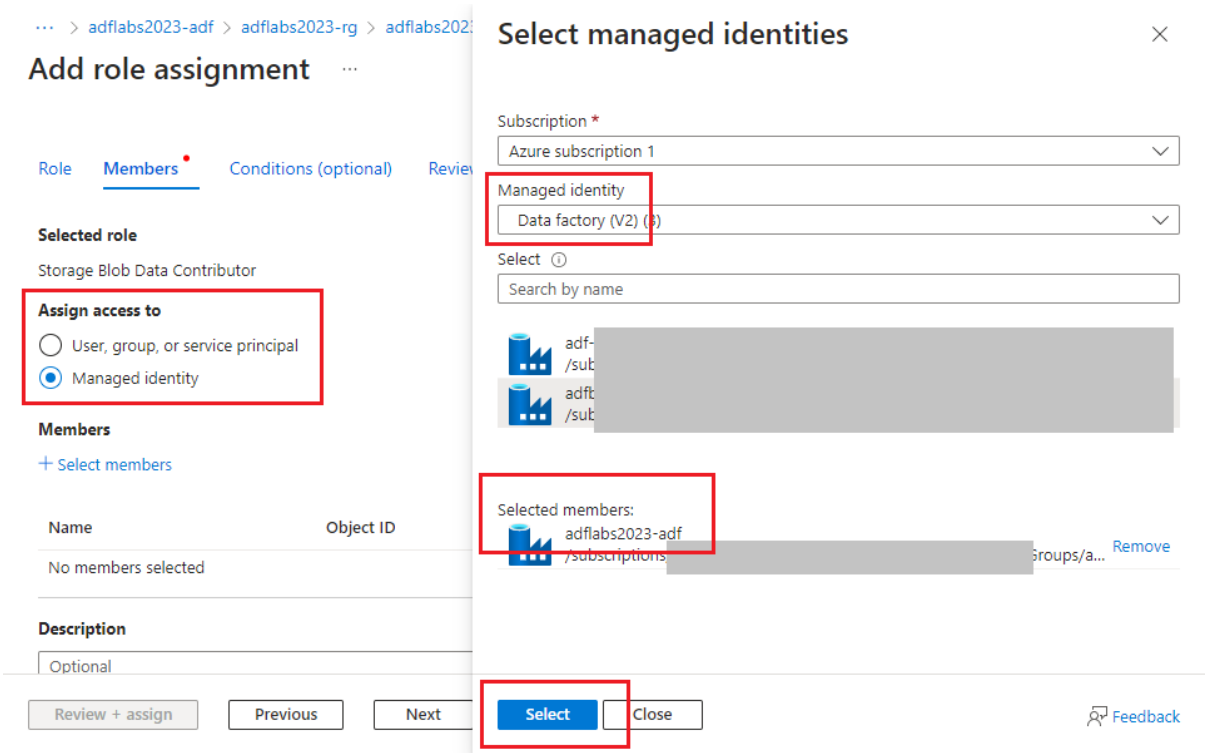
3. In the "Grant access to this resource" tile, click the "Add role assignment" button.

4. In the "Role" tab of the "Add role assignment" blade, choose the "Storage Blob Data Contributor" role – this authorises read, write and delete access in your data lake. Click "Next".





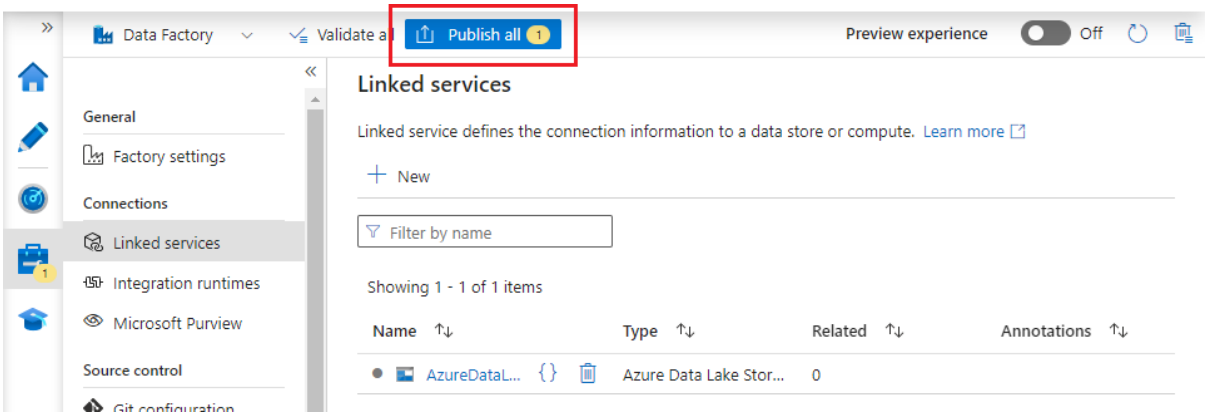
5. On the "Members" tab, specify that you wish to assign access to a managed identity, then click "+ Select members". In the "Select managed identities" flyout, choose "Data factory (V2)" from the "Managed identity" dropdown, then choose your factory from the displayed list of identities. Click "Select" to close the flyout.



7. Back in the "Add role assignment" blade, click "Review + assign", then "Review + assign" again to add the managed identity to the blob contributor role for your data lake.

8. Return to ADF Studio's Management Hub, then click on your data lake linked service to re-open the editing flyout. 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 linked service flyout.

9. When a data factory is not connected to a Git repository, the only way to save your changes is to publish them – click "Publish all" in the top menu bar to do so.



Recap

In Lab 1 you have:

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