Data Movement: From On-Premises to Data Lake Store

Updated on: 1/13/2017

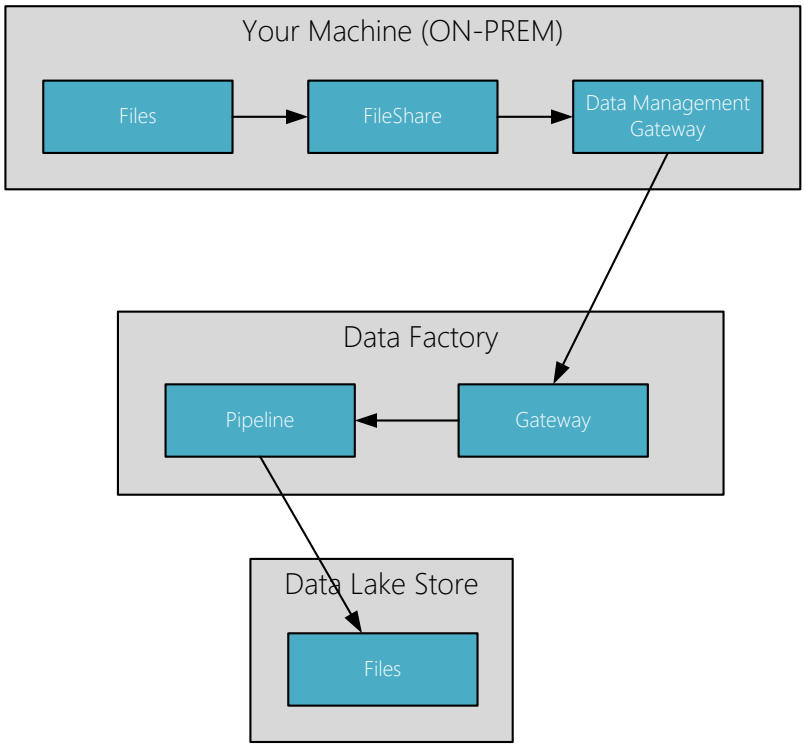
Introduction

As customer adoption of Azure Data Lake Store increases, there will be many who will have data stored on on-premises solutions These customers will want that on-premises data to be moved to Azure Data Lake Store in a one-time or recurring fashion. **Azure Data Factory** provides you an easy way to achieve this data copy in an intuitive, scalable, recurring fashion with appropriate data transformation.

In this lab, learn the basics of how to copy data from file shares stored on-premises to ADLS in a one-time fashion using Azure Data Factory and Data Management Gateway.

Overview of What we will build

The diagram below shows the components that are involved in getting this lab done. In summary, you will start with files stored on-premises (on your own machine) and copy them to Azure Data Lake Store via **Azure Data Factory** and **Data Management Gateway** in an form for further analysis. An interesting tweak in this example is that the input files are compressed .gz (GZip) files and they will be transferred in that format, yet when then arrive on the Data Lake store they will be automatically unzipped.



Prerequisites

- Machine running the latest version of Windows.
- To be logged-in to the machine using your domain credentials
- Access to an Azure subscription that you can create resources in.
- Access to an ADLA Account that you can submit jobs to.
- Access to an ADLS Account that you can submit jobs to.

Things to Keep in Mind

- File and folder paths in Azure Data Lake Store are case-sensitive
- If you are logging in win Microsoft credentials, please go to http://aka.ms/AzurePortalProd to get the same version of the portal that customers see.
- You can use any browser, but for the purposes of this lab using Microsoft Edge or Microsoft Internet Explorer will be simpler since they natively support the ClickOnce feature

Step 1: Get the source files ready on your machine

- Create a folder at the root of a dive called **C:\ADLSTutorial**.
    - Right click on the **ADLSTutorial** folder, then select **Share With** > **Specific People**.

    - Choose your own user account
    - click **Share**.
- Inside that shared folder, create a folder called **Source** that will contain the files to be copied.
- Download the SmallDataset.zip file from here
- Extract the ZIP. It will contain two .gz files.
- Put the contents of the zip into the Source folder.

Your folder structure should look like

```
C:\
    ADLSTutorial\
        Souce\
            commitssmall.csv.gz
            userssmall.csv.gz
```

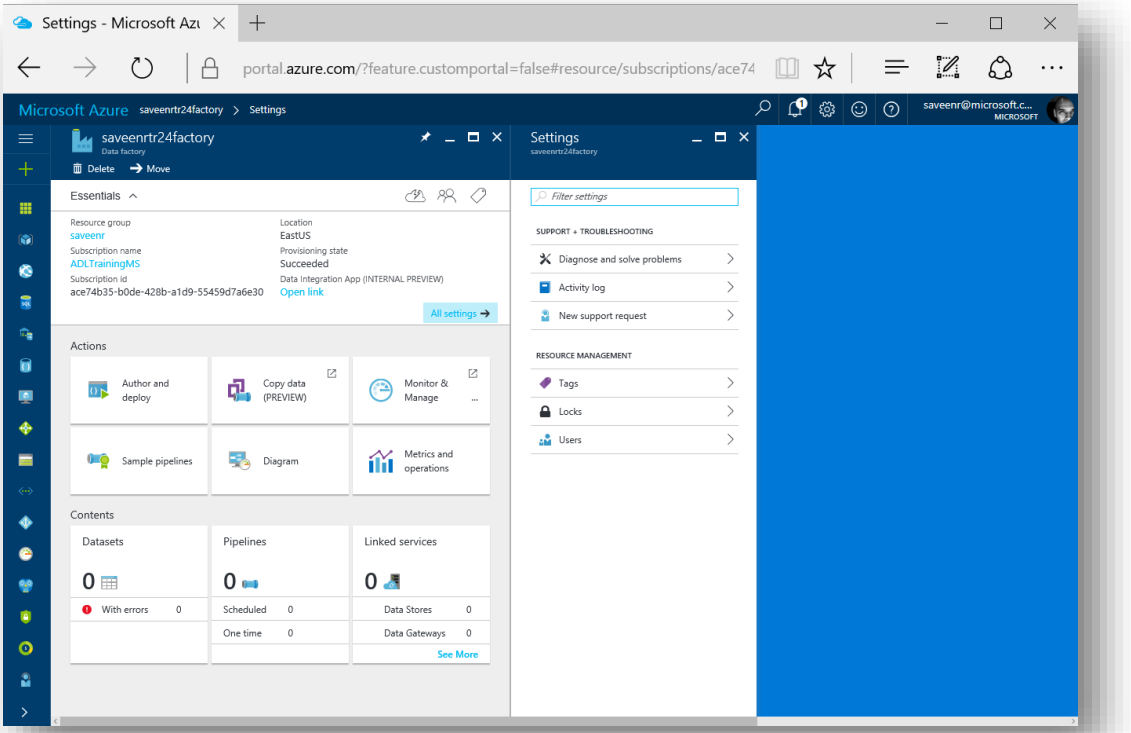
Step 2: Create the destination folder in ADLS

- Log into the Azure portal http://portal.azure.com
- Open your ADLS Account and create a folder called **/TR24**

Step 3: Create the Azure Data Factory

NOTE: The name of the Azure data factory must be globally unique. Try using something like "ADLSTutorial-<username>"

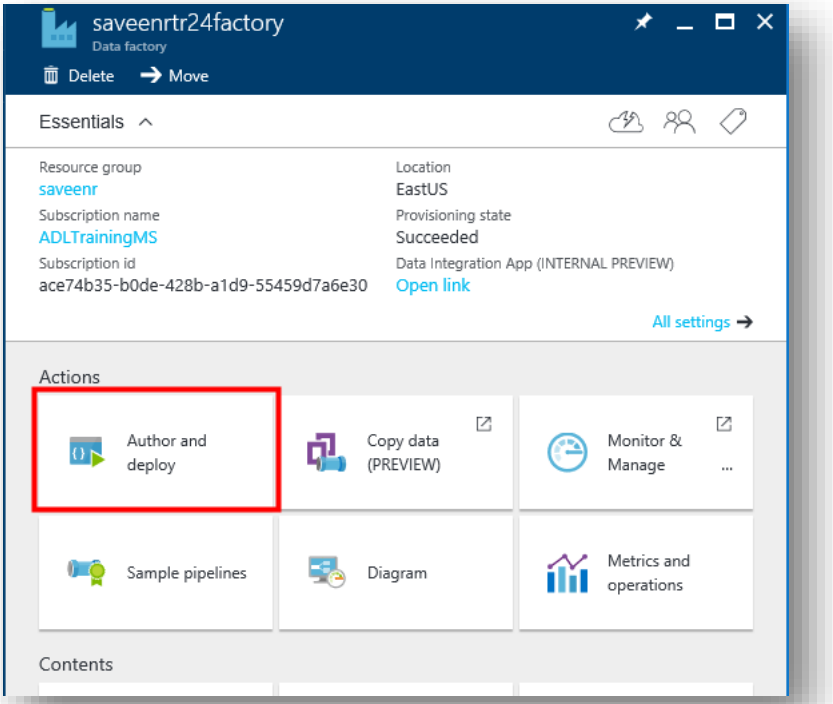
- Log into the Azure portal http://portal.azure.com
- click **New** > **Intelligence + Analytics** > **Data Factory** on the **Data + Analytics** blade. The **New data factory** blade will appear.
- In the **New data factory** blade:
    - Enter **ADLSTutorial-<username>** for the **name**.
    - For **Resource Group**, use whatever you want
    - For **Location**, leave with the default selection.
- Enable **Pin to dashboard** checked box.
- Click **Create**
- Once creation is complete, you will see the **Data Factory** blade as shown below:



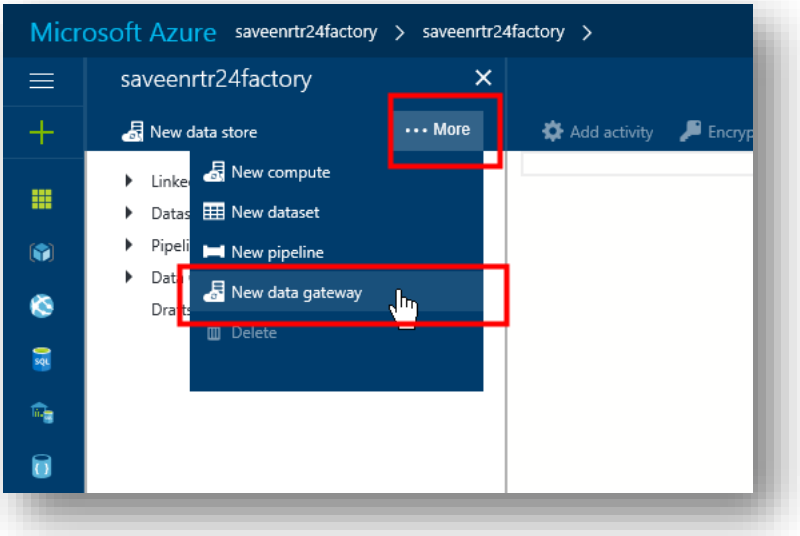
Step 4: Create a Data Management Gateway

IMPORTANT: if Data Management Gateway is already installed on your machine, uninstall it

- Navigate to **Data Factory** that you created
- Click **Author and deploy** tile to launch the **Editor** for the data factory.



- In the Data Factory Editor, click **... (ellipsis)** on the toolbar
- Click **New data gateway**

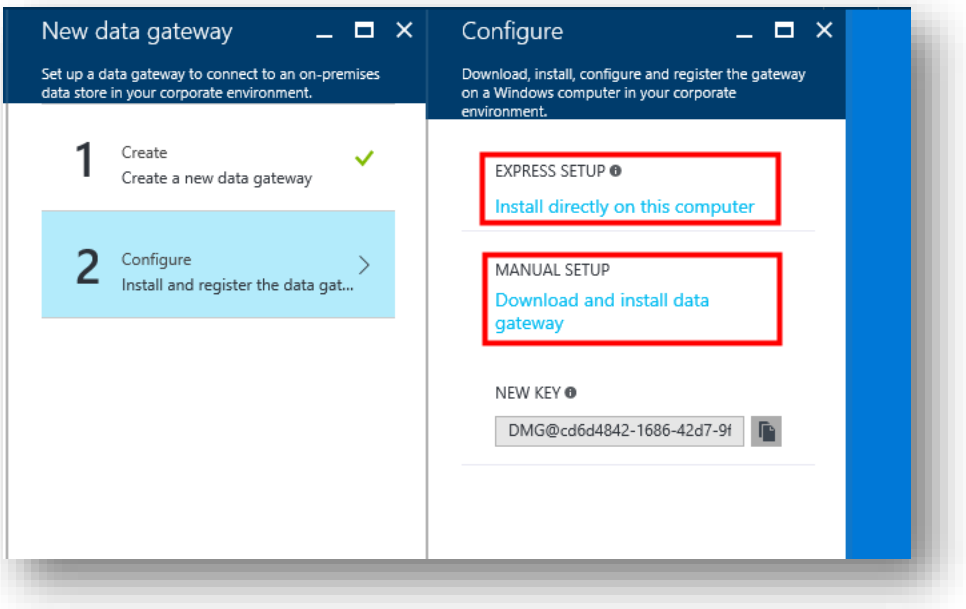


- The **New data gateway** blade will appear
- In the **Create** step, for the **Data gateway name**, enter **ADLSGateway-<username>-TR24**
- Click **OK**
- In the **Configure** step, you'll have two ways of installing DMG: EXPRESS and MANUAL

a. EXPRESS: Under **EXPRESS SETUP** click **Install directly on this computer**. This will download the installation package for the gateway, install, configure, and register the gateway on the computer using ClickOnce. This is the easiest way of installing DMG but please verify you have met the requirements for ClickOnce as defined in the appendix of this document.

b. MANUAL: Under MANUAL SETUP click on **Download and install data gateway.**

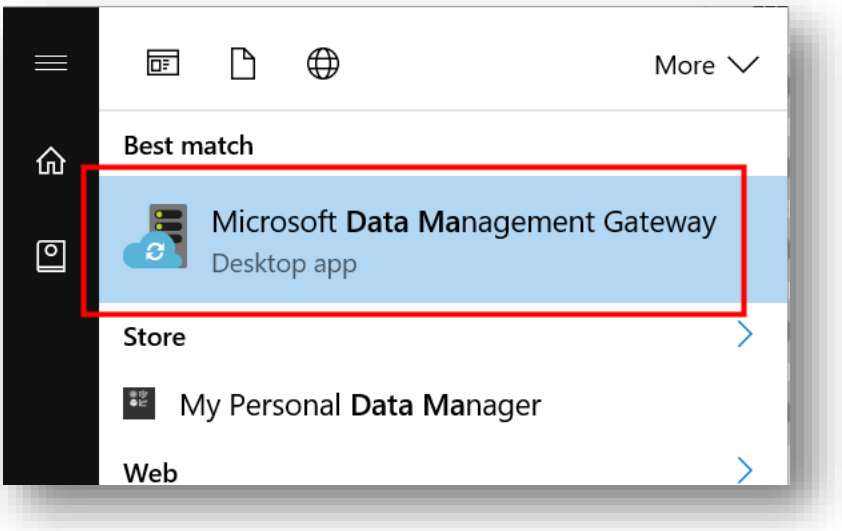
- NOTE: You must be an administrator on the local computer to install and configure the Data Management Gateway successfully. You can add additional users to the Data Management Gateway Users local Windows group. The members of this group will be able to use the Data Management Gateway Configuration Manager tool to configure the gateway.



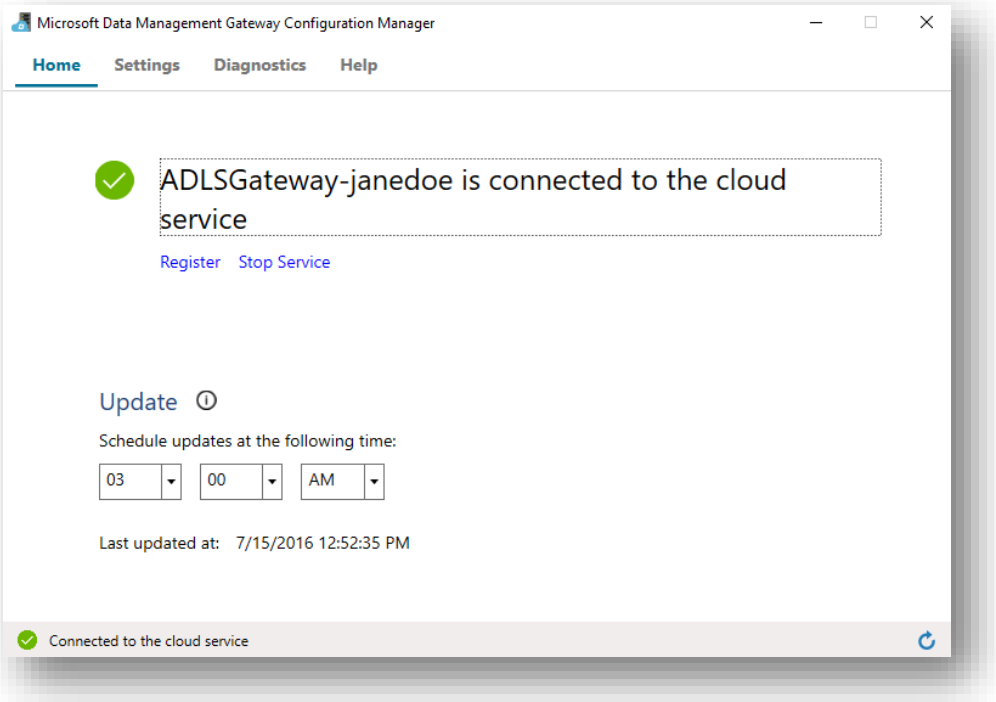
- To verify that DMG hs been installed, check that the executable **ConfigManager.exe** exists in the folder: **C:\Program Files\Microsoft Data Management Gateway\1.0\Shared**.

NOTE: You must be an administrator on the local computer to install and configure the Data Management Gateway successfully. You can add additional users to the Data Management Gateway Users local Windows group. The members of this group will be able to use the Data Management Gateway Configuration Manager tool to configure the gateway.

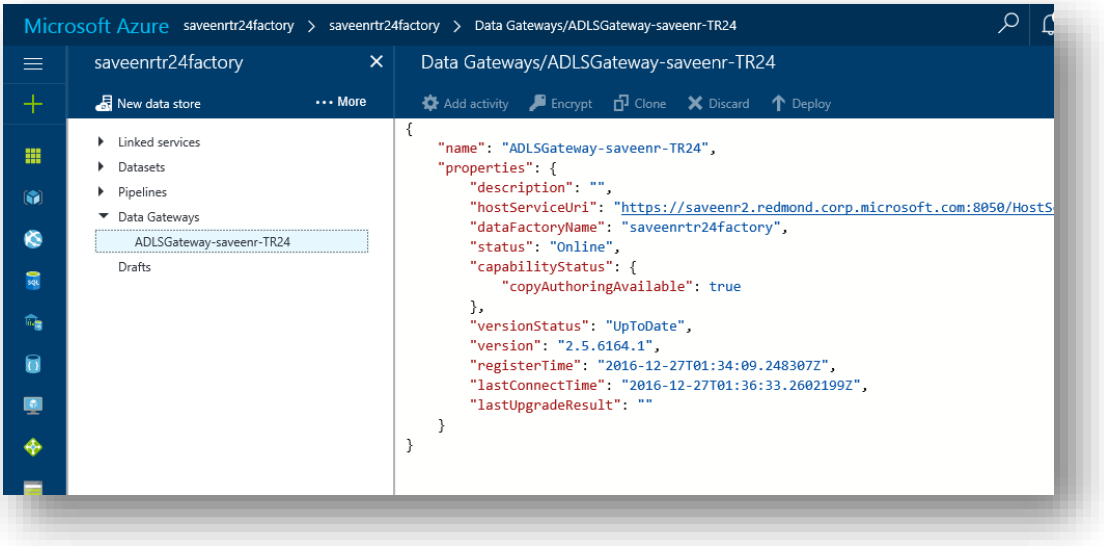
- Wait for a couple of minutes and ensure that Data Management Gateway is installed correctly. Once you see it in the start menu, click on it



- When DMG Starts

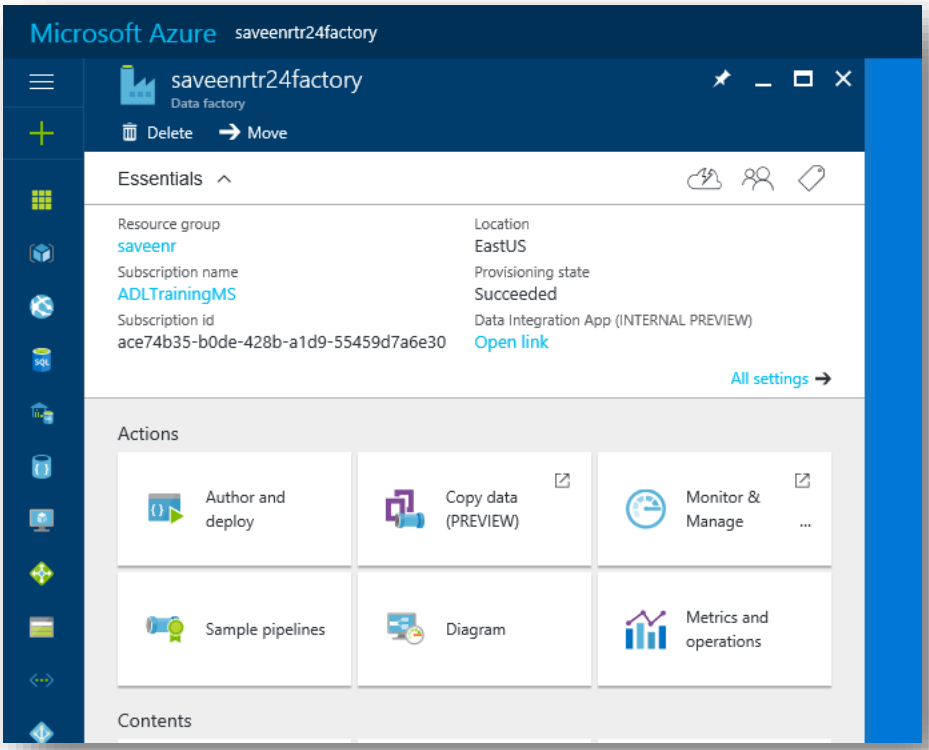


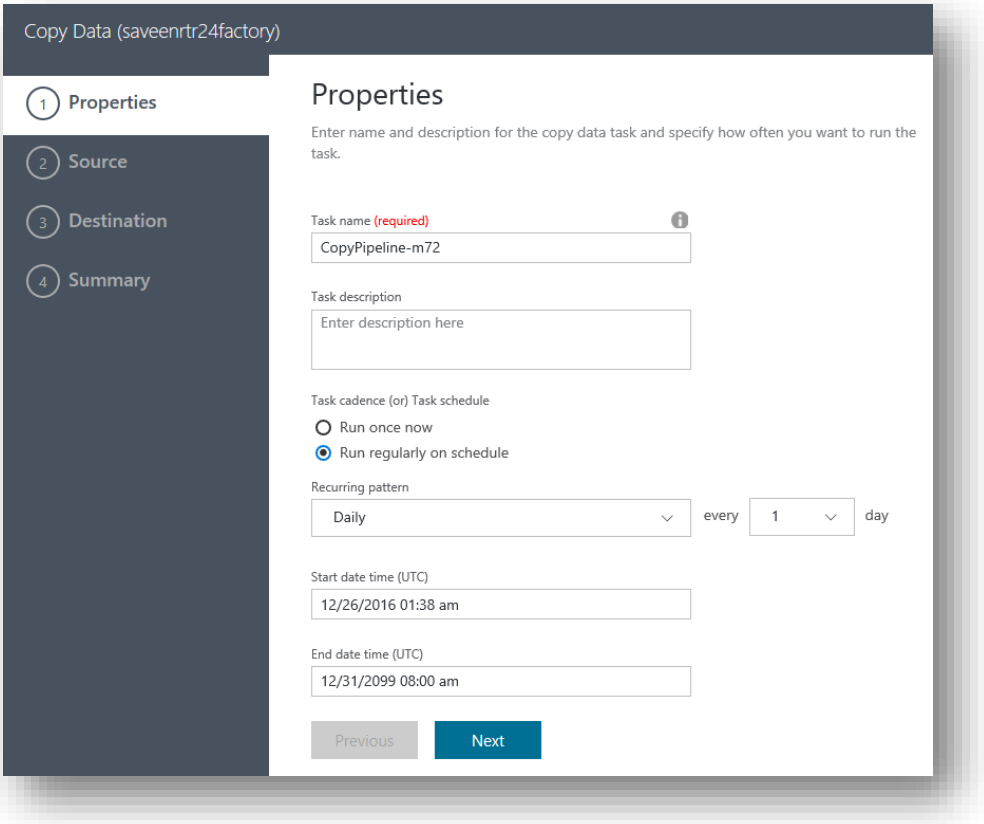
- Wait until the you see the message that it is **connected to the cloud service** as shown above.
- Go back to the Azure Portal, click **OK** on the **Configure** blade and then on the **New data gateway** blade.
- You should see **ADLSGateway-<your alias>** under **Data Gateways** in the tree view on the left. If you click on it, you should see the associated JSON.



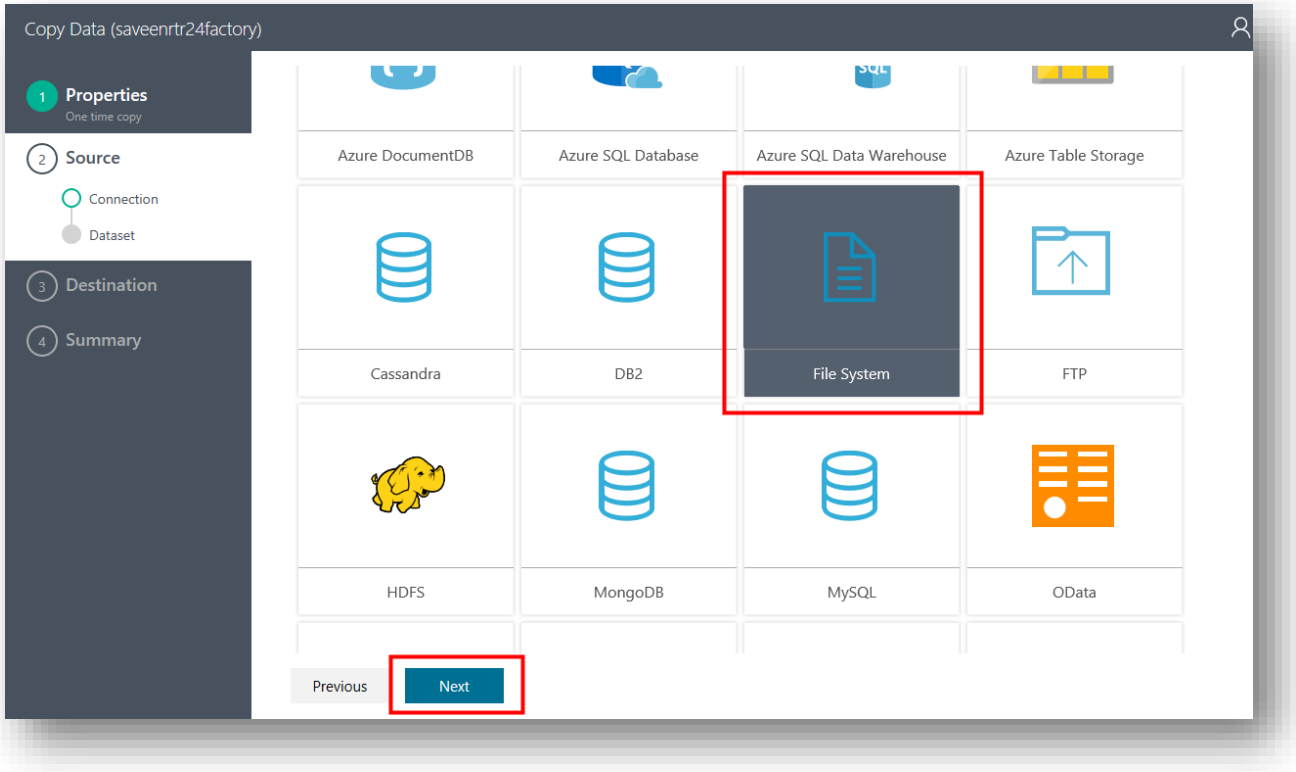
Step 5: Create and run a pipeline using Copy Wizard

- Go to the Azure Portal
- Navigate to the ADF factory that you created
- Click **Copy data**... The **COPY DATA – Properties** page will appear.

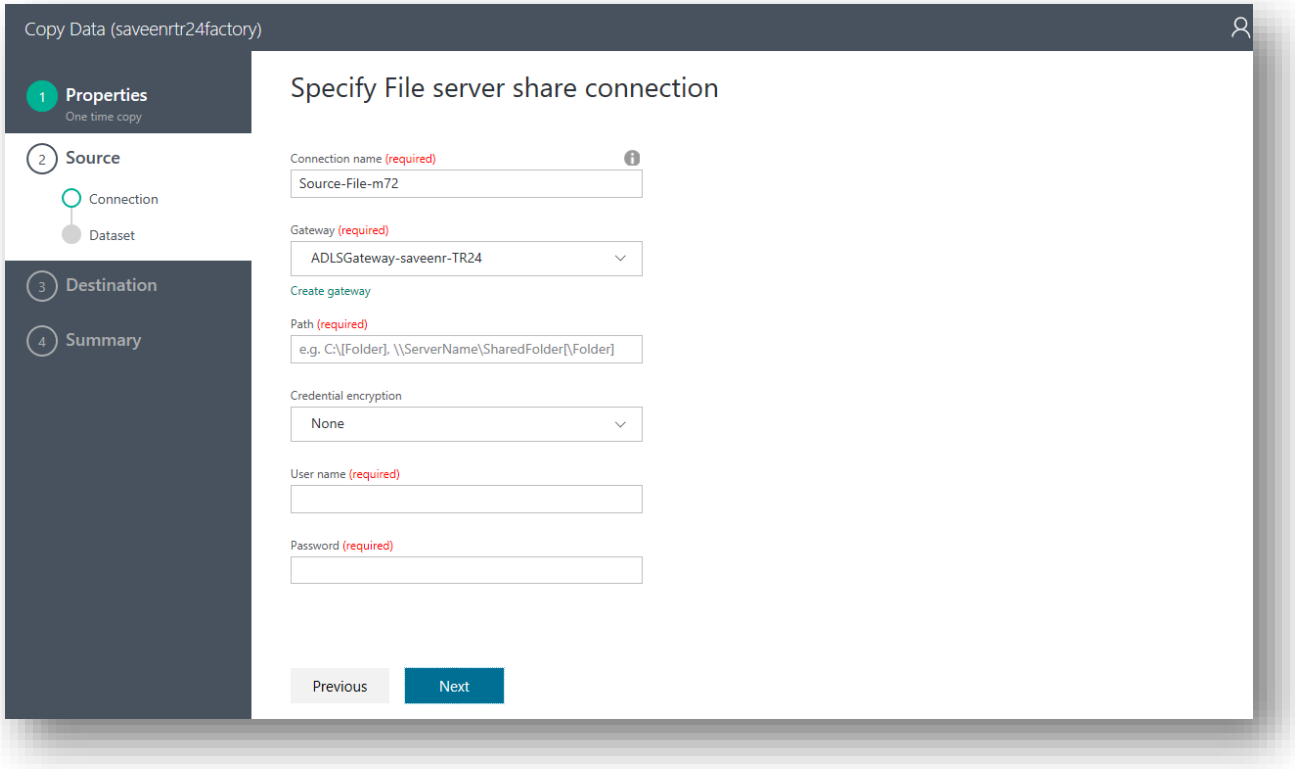

Microsoft Azure
saveenrtr24factory
saveenrtr24factory
Data factory
Delete
Move
Essentials
Resource group
saveenr
Location
EastUS
Subscription name
ADLTrainingMS
Provisioning state
Succeeded
Subscription id
ace74b35-b0de-428b-a1d9-55459d7a6e30
Data Integration App (INTERNAL PREVIEW)
Open link
All settings
Actions
Author and deploy
Copy data (PREVIEW)
Monitor & Manage
Sample pipelines
Diagram
Metrics and operations
Contents



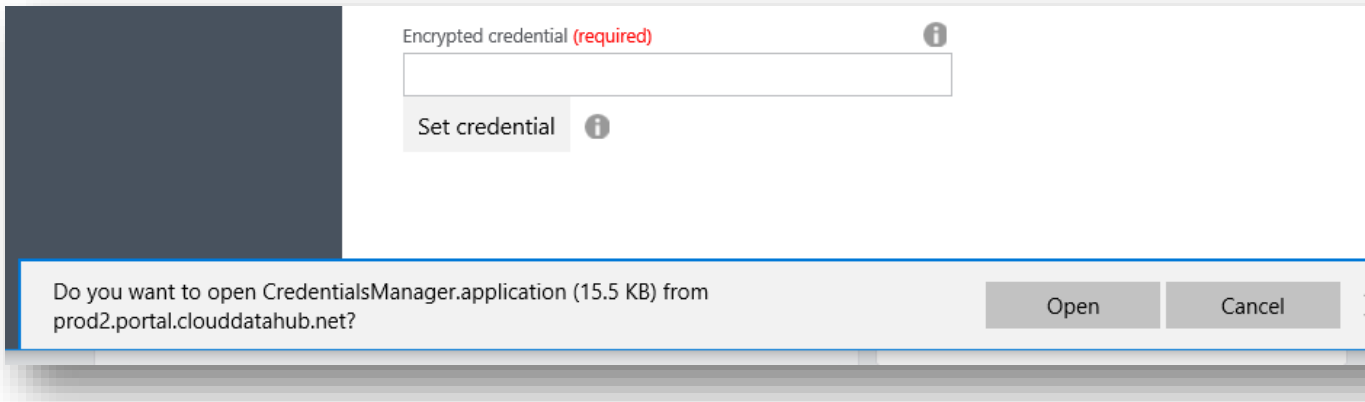
- Task name – Enter a name that you will be familiar with, otherwise note down the name that is prefilled e.g. CopyPipeline-l17 here. Note down this name on a piece of paper, you will need it later to monitor this pipeline.
- Click on **Run once now**. You can keep all the other values as the default ones.
- Click on **Next**.
- You will be taken to the **COPY DATA – Source data store** page.
- On the **COPY DATA – Source data store** page, click on **File System** on the Connect to a Data Store screen.
- Click on **Next**.



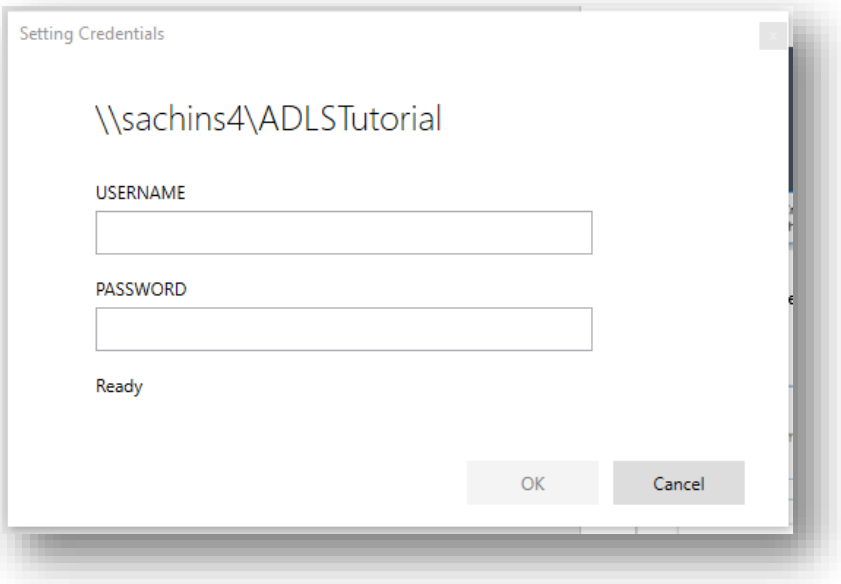
- You will be taken to the **COPY DATA – Specify File server share connection** page.



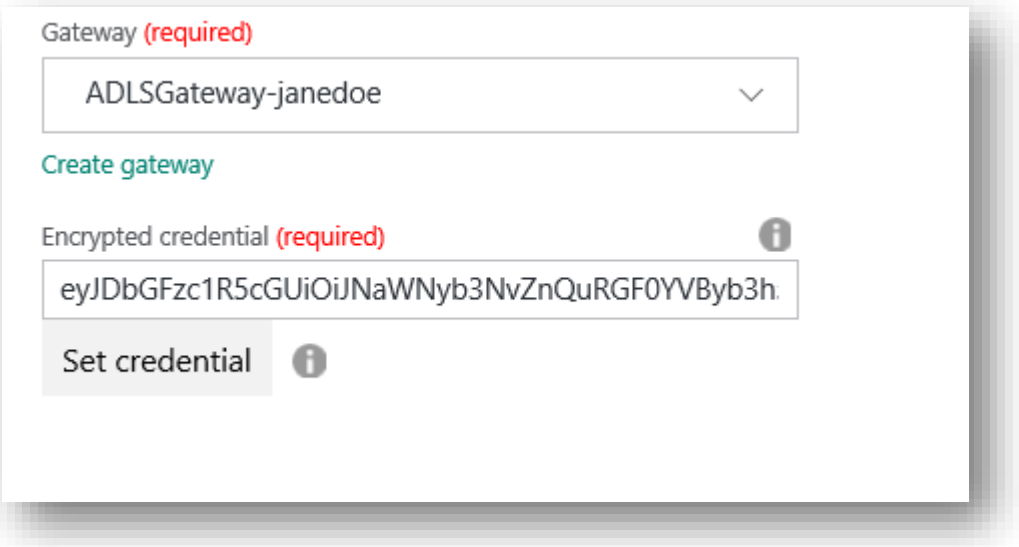
- On the **COPY DATA - Specify File server share connection** page, as shown above, enter following values for the field.
- **Linked service name** – Leave it as is.
- **Path** – Full path for the share you created above i.e. \\janedoelaptop\ADLSTutorial
- **Credential encryption** – set to **Use credential manager**.
- **Gateway –** It will be populated automatically with the one you created above i.e. ADLSGateway-<username>
- Click on **Set credential**.
- Allow the CredentialManager.application to be installed.



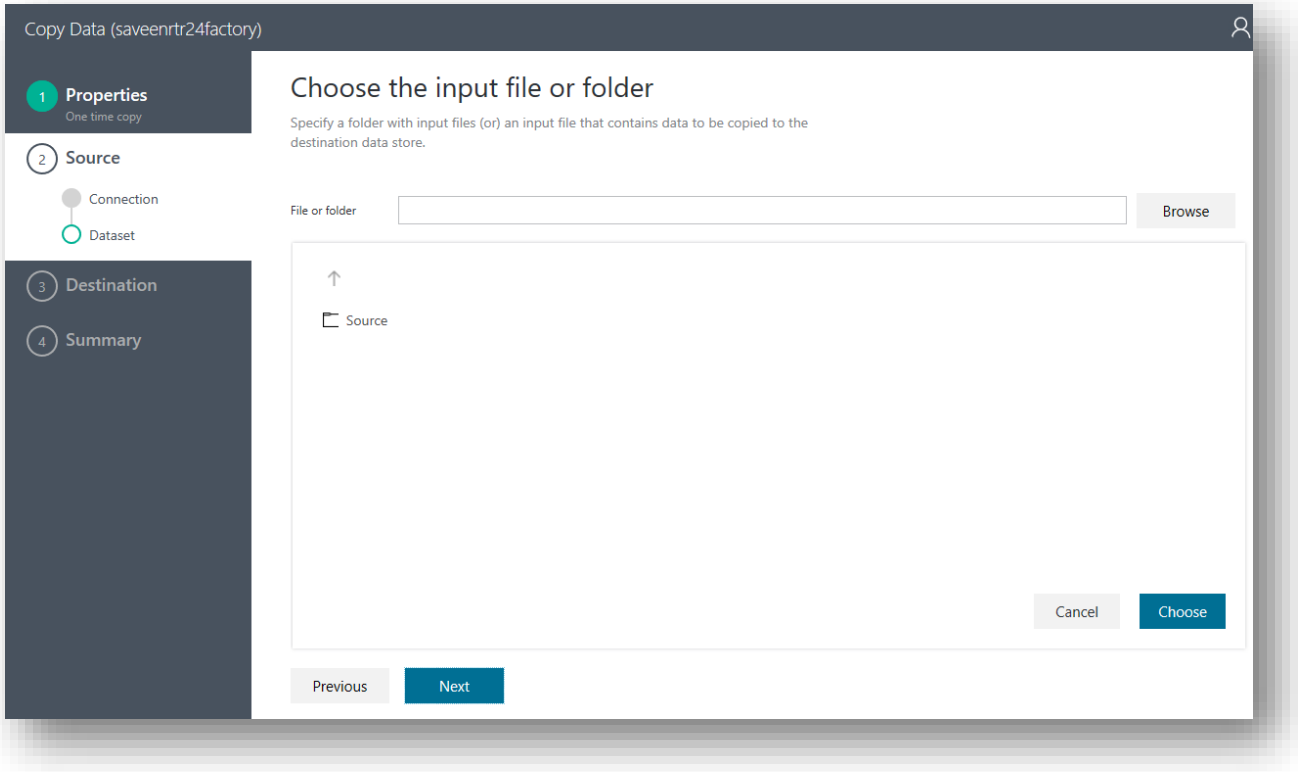
- You will be shown the **Setting Credentials** screen, where you will enter the credentials that allow access to the share. Since you are using a domain joined redmond machine, your Microsoft credentials should work i.e. username=<domain>\<username> and appropriate password.



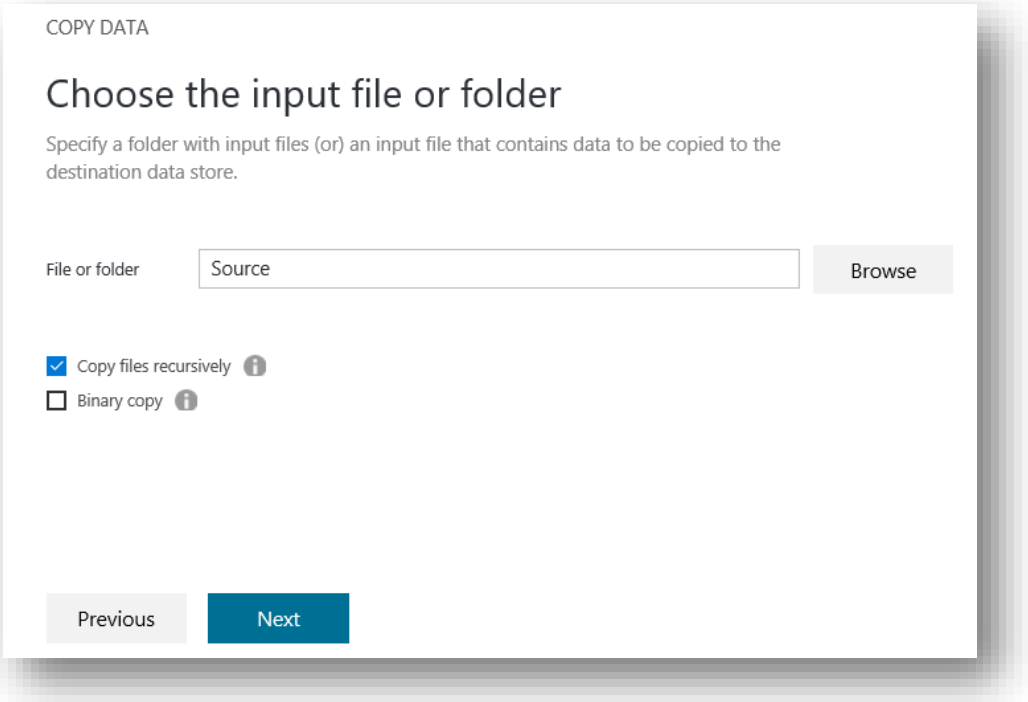
- Click on **OK** after entering the credentials.
- If you entered your credentials correctly, you should see the **Encrypted credential** field appropriately filled.



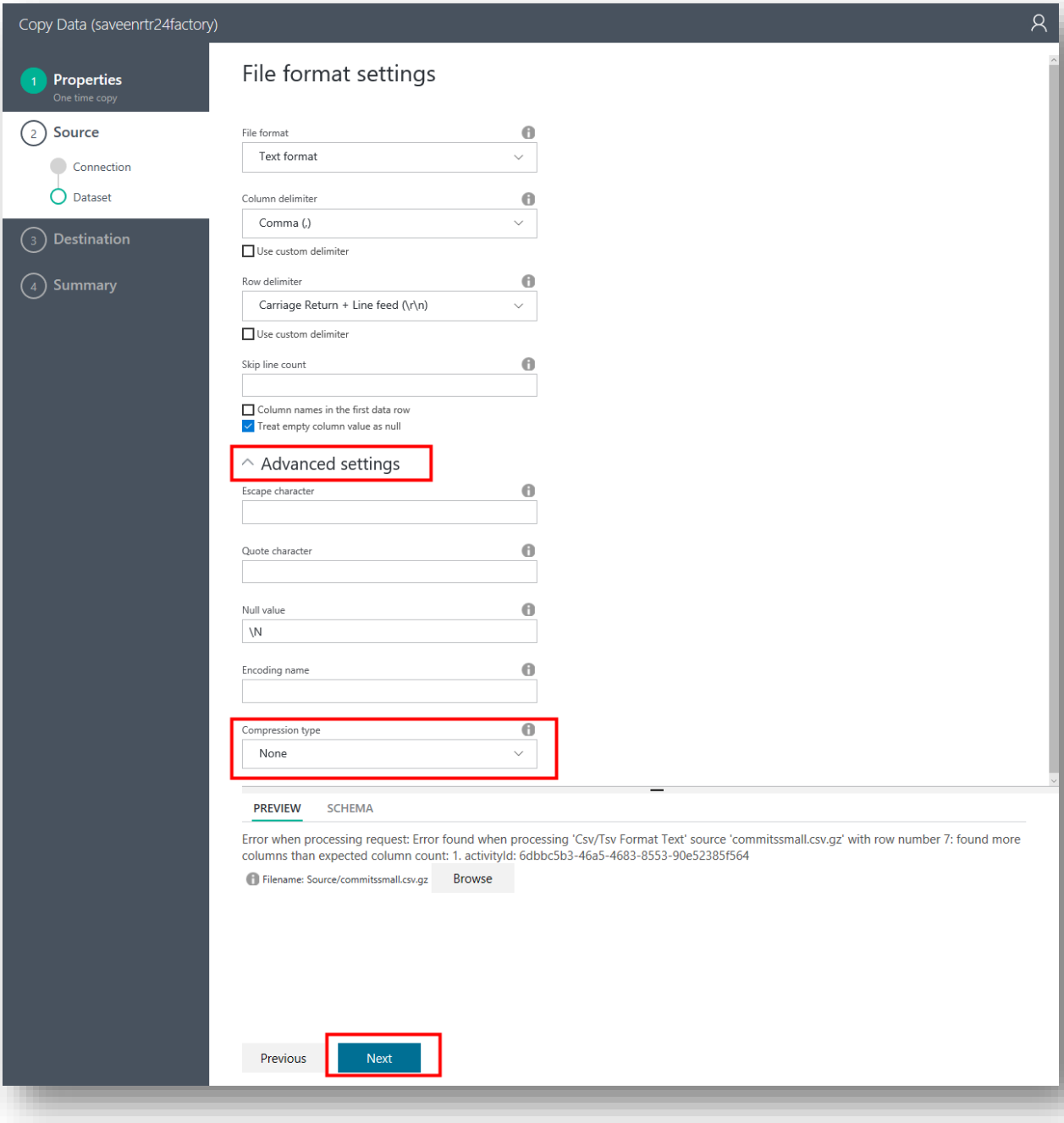
- Click on **Next**. You will be taken to the **COPY DATA – Choose the input file or folder** page.



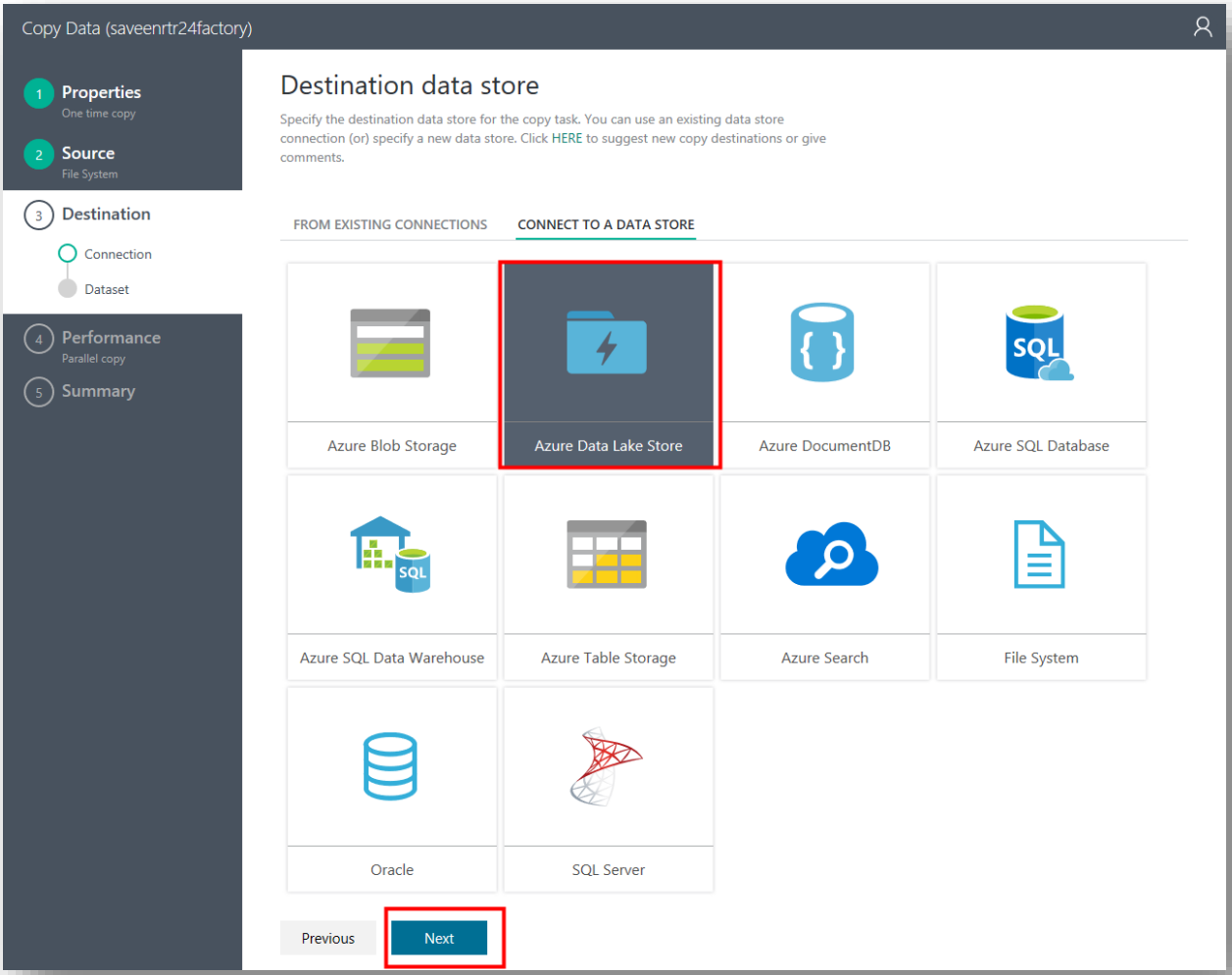
- On the **Copy Data – Choose the input file or folder** page, as shown above, click on the **Source** folder that contains the gzip files and then click on **Choose**. On the same page, you will then be provided additional options you can set for the copy.



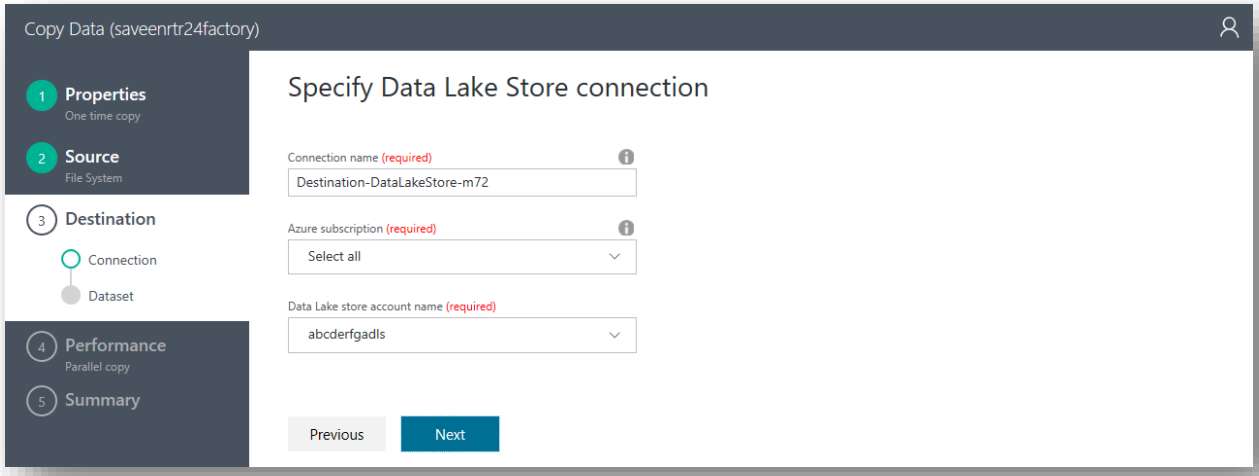
- For this tutorial, click on **Copy files recursively** option. This is to ensure that the all the data in the folder is copied over.
- Click on **Next**.
- You will be taken to the **COPY DATA – File Format settings** page.



- On the **COPY DATA – File format settings** page as shown above, expand the **Advanced settings** section. Scroll down and select the **Compression type** to **GZip**. You should see a preview of the data in the first source file as shown above.
- Click on **Next**.
- You will be taken to the **COPY DATA – Destination data store** page.

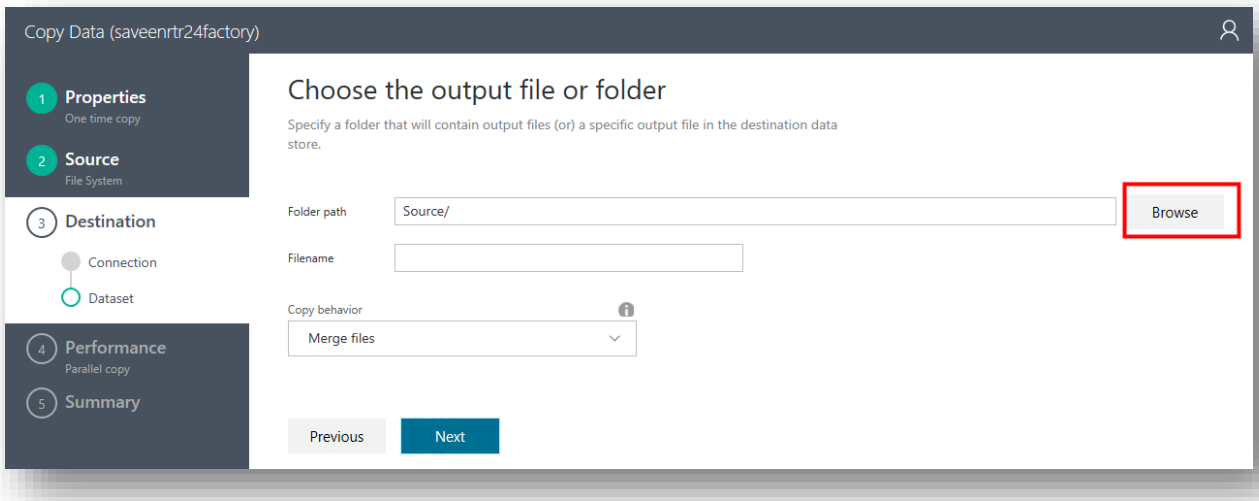


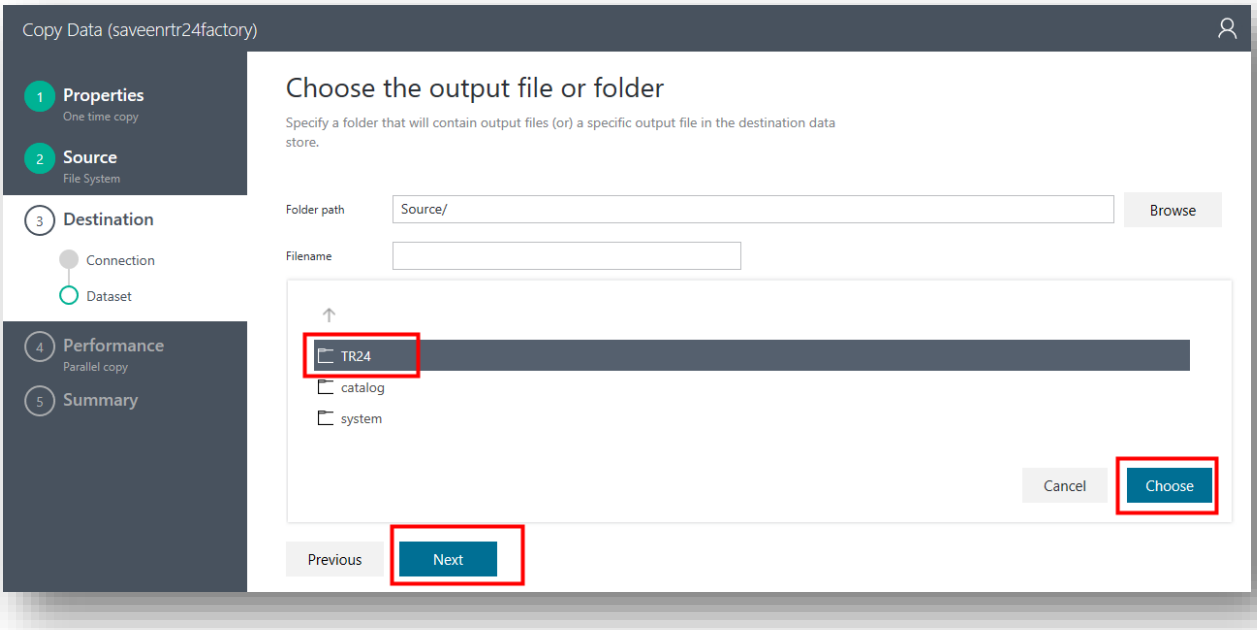
- Here you choose the **Azure Data Lake** as the item.
- Click on **Next**.
- You will be taken to the **COPY DATA – Specify Data Lake Store connection** page.



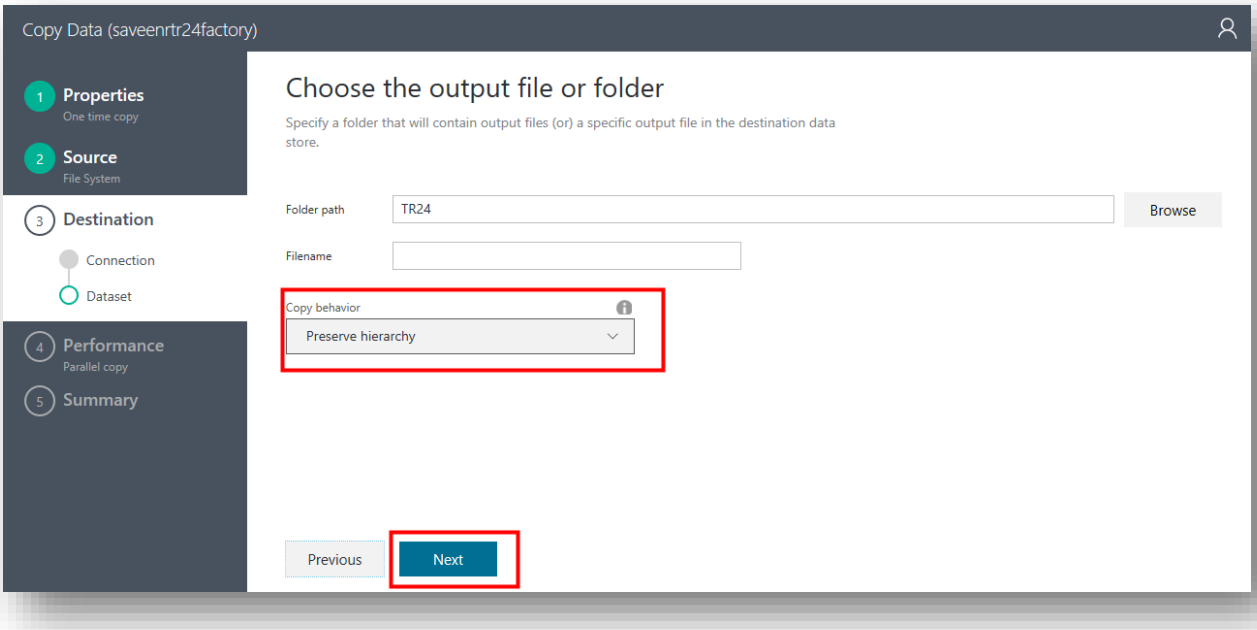
- On the **COPY DATA – Specify Data Lake Store connection** page, as shown above, enter the following fields:
- **Linked service name** – Leave as in.
- **Data Lake store account name** - Pull down and select the relevant entry.
- Click on **Next**.
- You will be taken to the **COPY DATA – Choose the output file or folder** page.

- On the **COPY DATA – Choose the output file or folder** page, shown above, first thing is to choose the **Folder path** where you will land the data. To do this, click on Browse and navigate to TR24/Users/janedoe/ADLSTutorial and select **Destination**. As you will recall, this is the directory that you created in step 2 earlier.

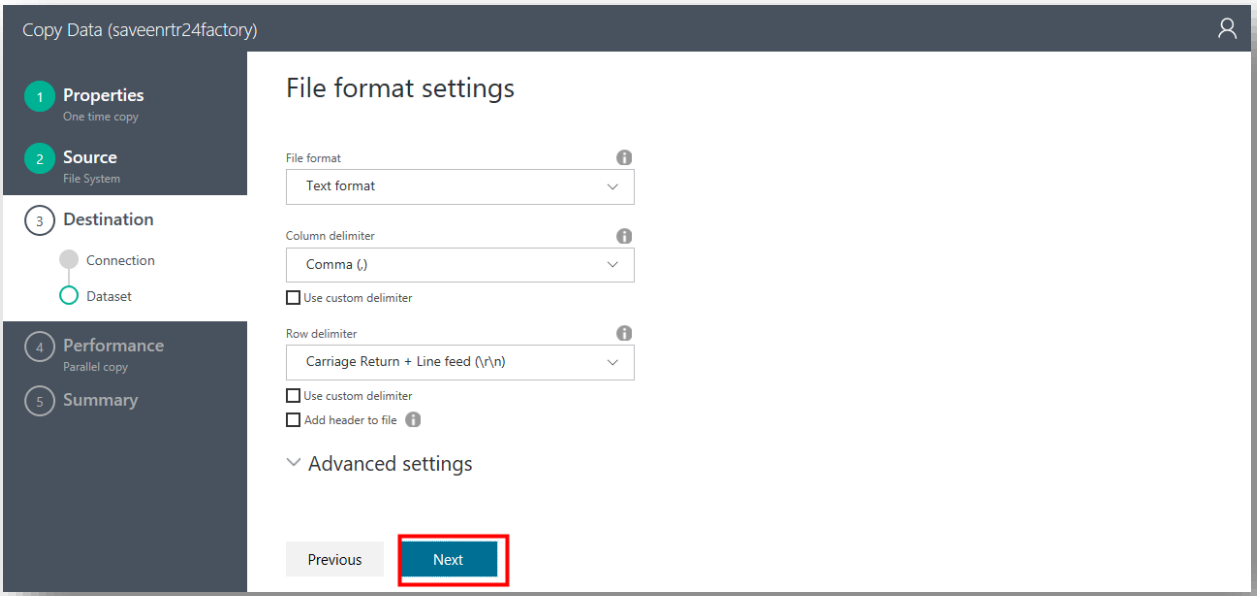




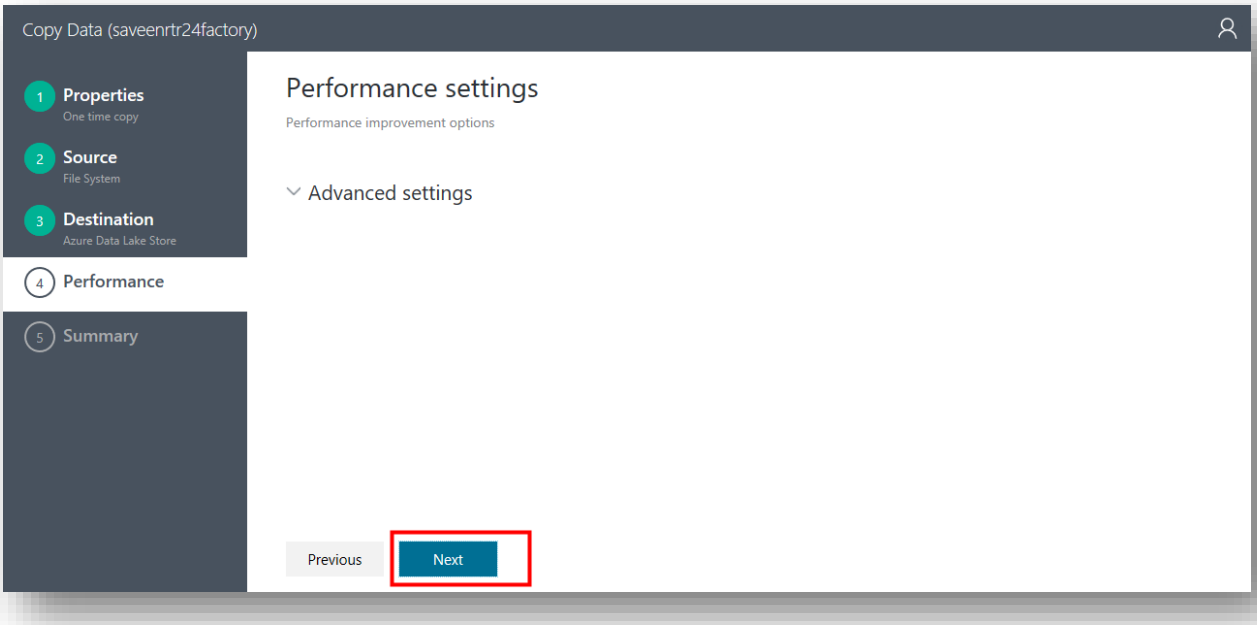
- Click on **Choose**. You will see the page below.



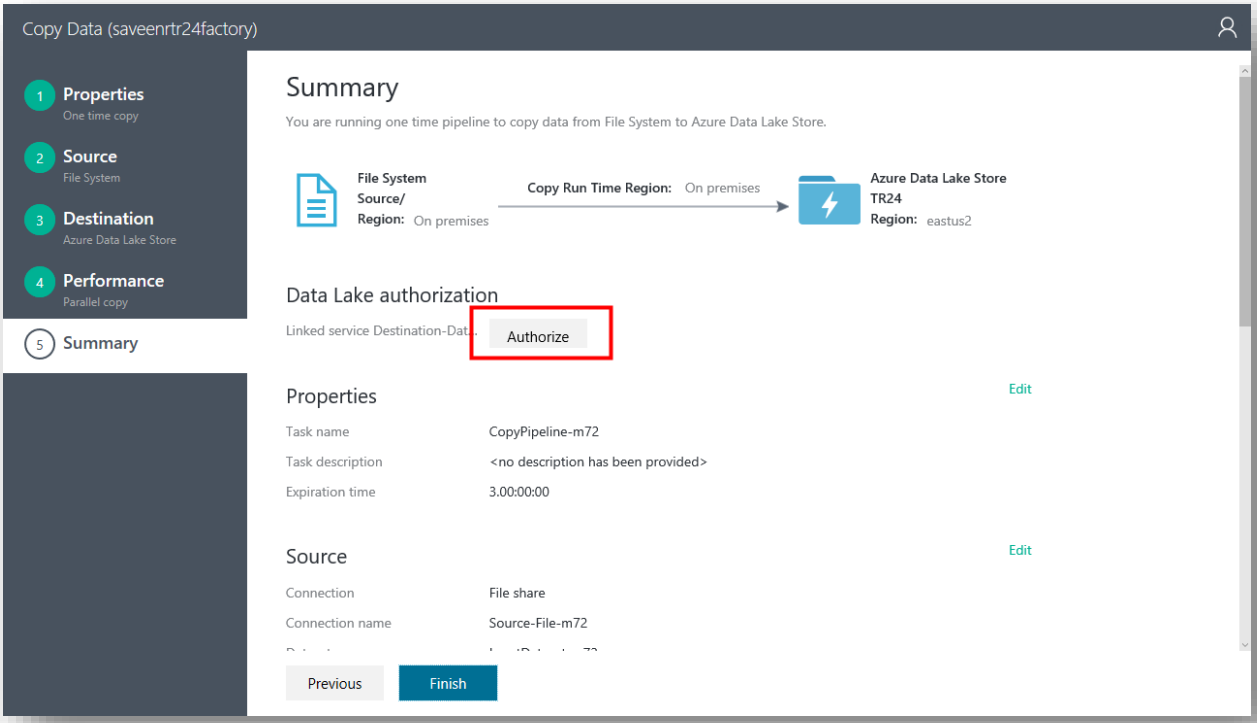
- Leave the **Filename** filed as is.
- In the **Copy behavior** pull down, select the "**Preserve hierarchy**" option.
- Click on **Next**. You will be taken to the **COPY DATA – File format settings** page.



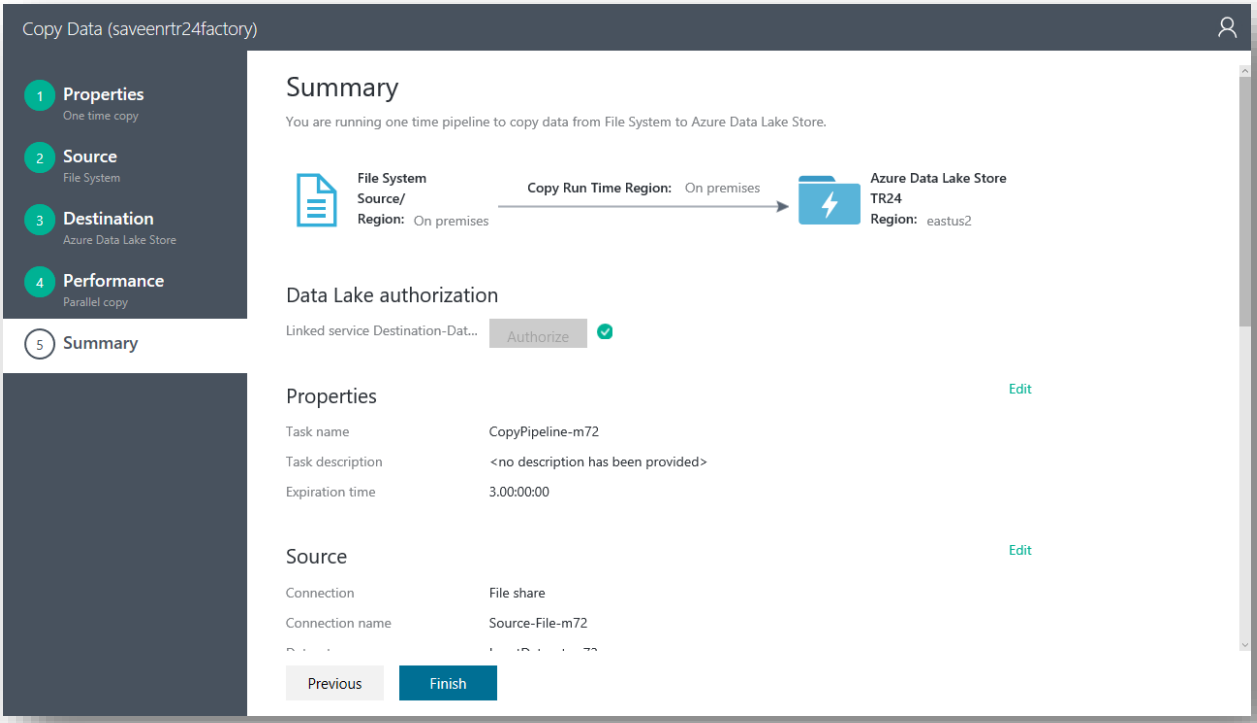
- On the **COPY DATA – File format settings** page, do not select anything. Just click on **Next**.



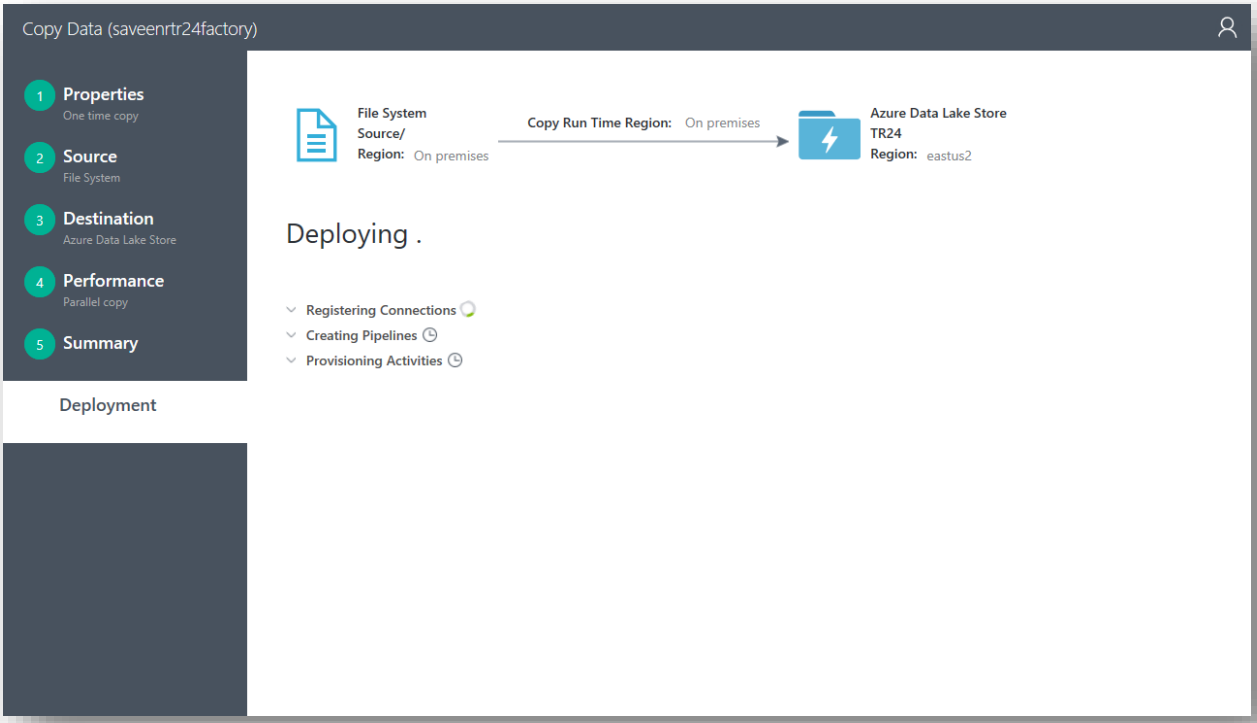
- You will be taken to the **COPY DATA – Summary** page.



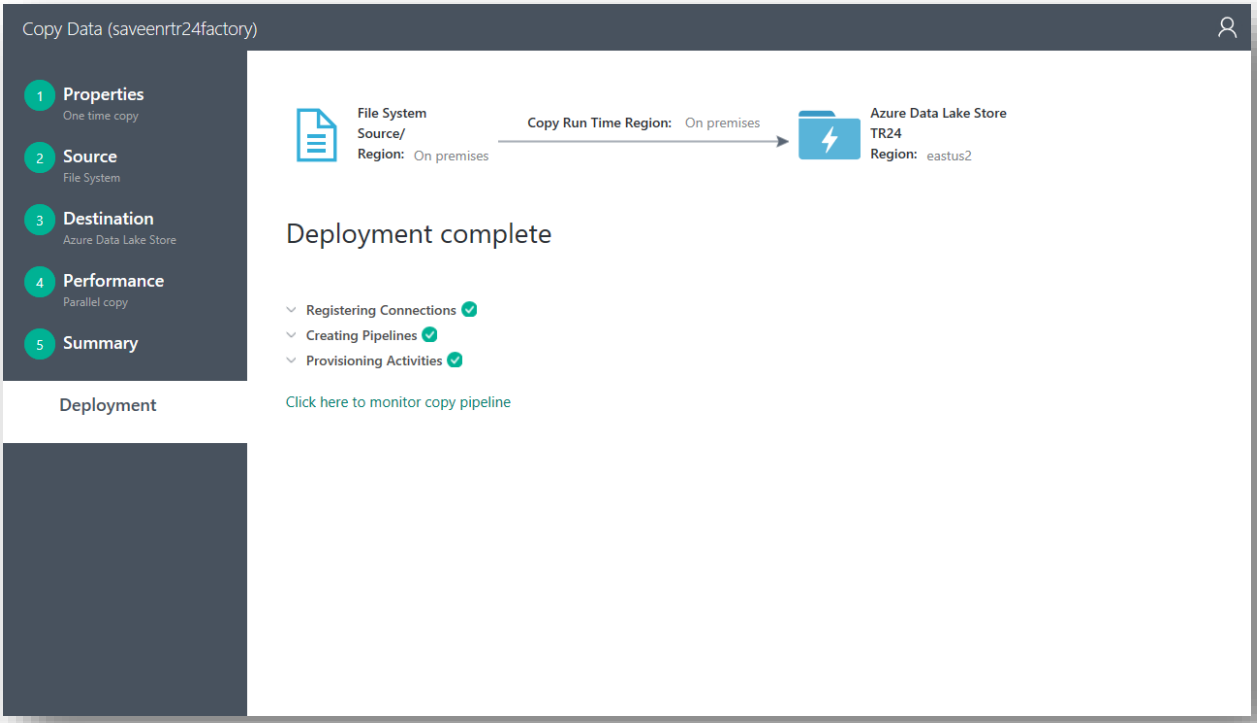
- On the **COPY DATA** - **Summary** page, as shown above, first click on the **Authorize** button in the **Data Lake authorization** section. This will authorize ADF access to ADLS. You will be prompted for credentials that have access to the ADLS account we are accessing.
- You should see a green checkmark once Authorization has completed successfully.



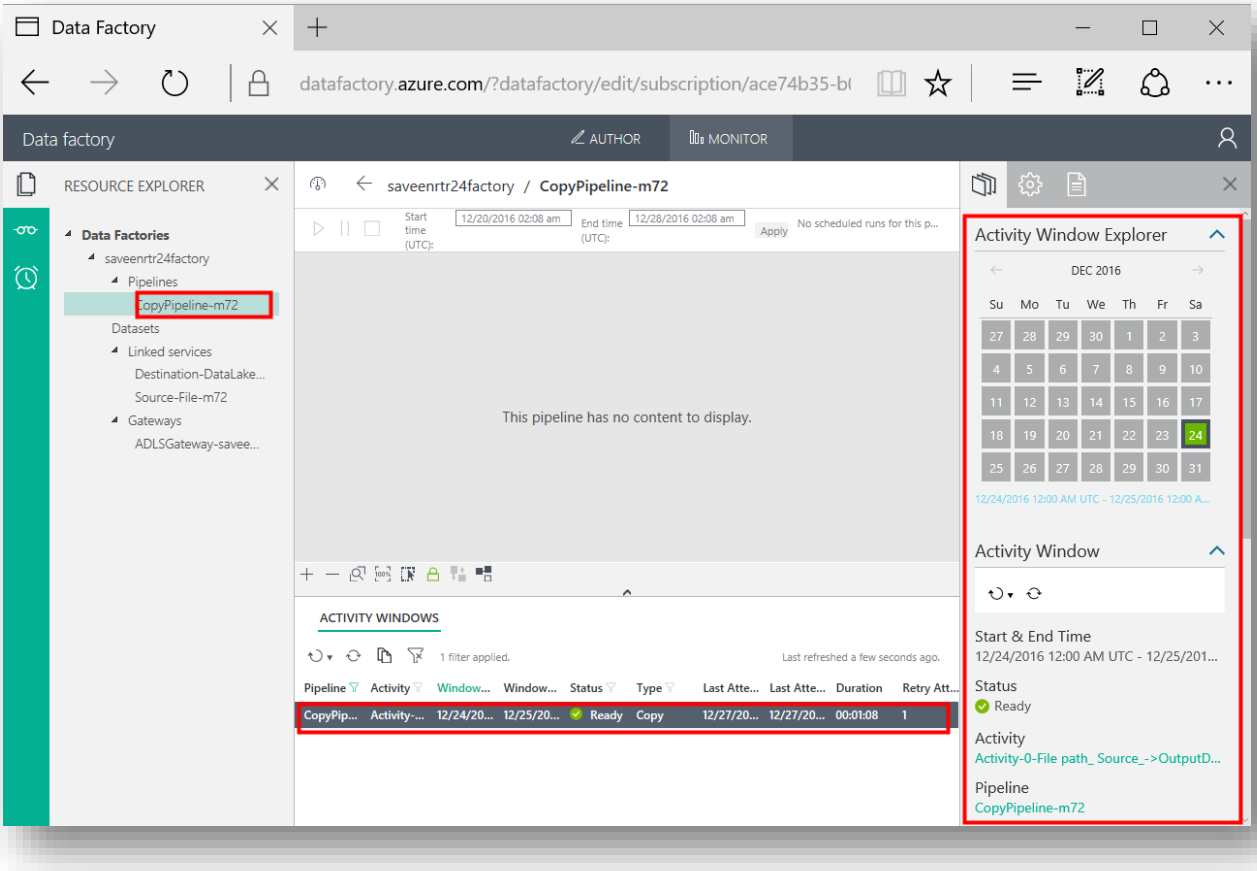
- Click on **Finish** and wait. You will be taken to the **COPY DATA - Deploying** page.



- In the **Deploying** screen where you will see **Registering Connections** and **Creating Pipelines**.
- Once there are green check marks next to those, you will see the title change to Deployment complete.

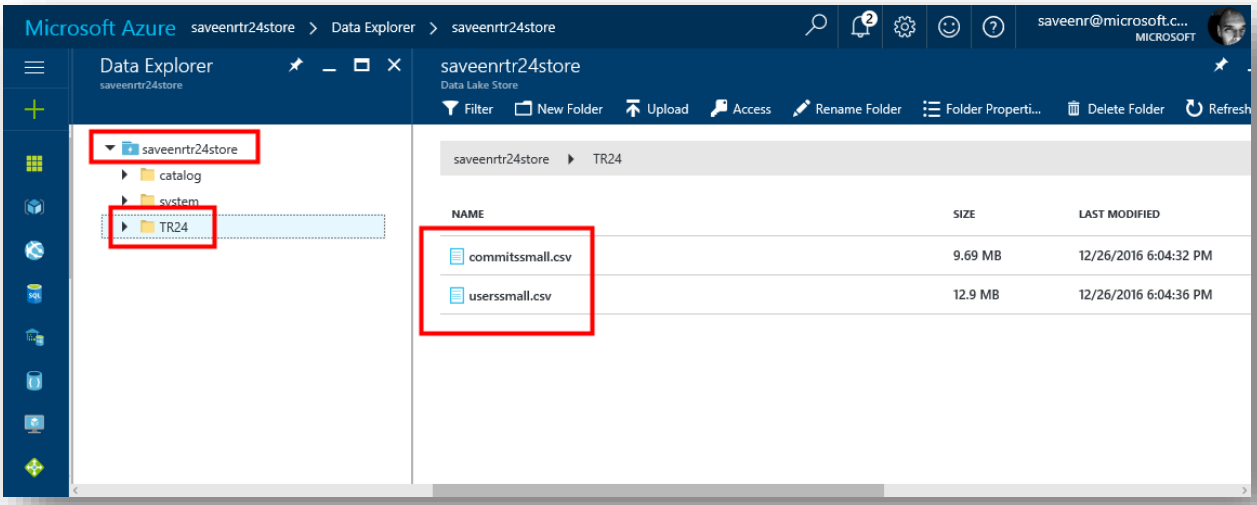


- At this point, recall the name of the Pipeline that you created earlier e.g. in this tutorial it was **CopyPipeline-l17**. Or you can find it be clicking on the down arrow next to **Creating Pipelines**.
- You will additionally see **Click here to monitor copy pipeline** link. Click on that link. You will be taken to the monitoring UI that monitors all your pipelines.



- Wait for the copy activity to be completed. Once you see it has succeeded in the right pane, go to the **MyStoreAccount** ADLS account via the Azure portal. Using Data explorer navigate to the TR24/users/janedoe/ADLSTutorial/Destination folder.

`



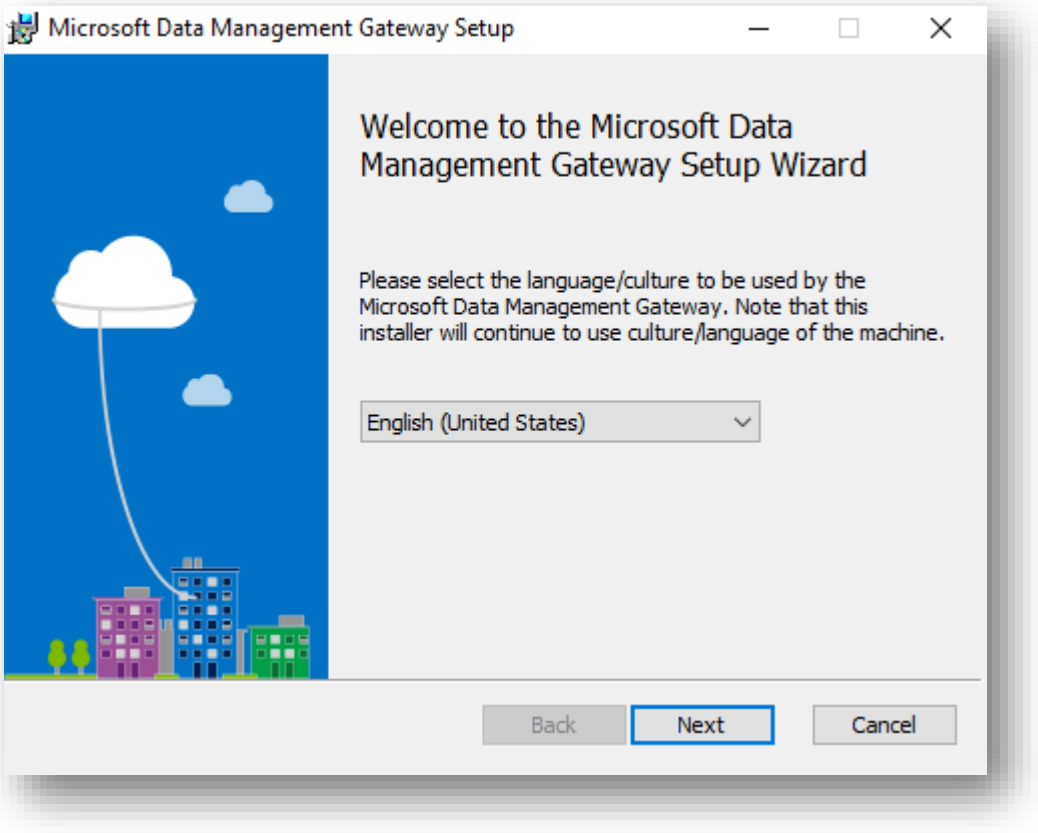
- You should see 2 files in the folder as shown above. Look at the **LAST MODIFIED** column and confirm that your file was recently copied over. Also confirm that the file does not have the "gz" suffix anymore. This shows that the gzip files were successfully unzipped and copied over from your machine.
- **Congratulations!** You have taken multiple gzip files stored in a folder from an on-premises location, and copied them over to Azure Data Lake Store in the cloud in an uncompressed form. You are now ready to process them using HDInsight, Azure Data Lake Analytics or an analytics tool of your choice.

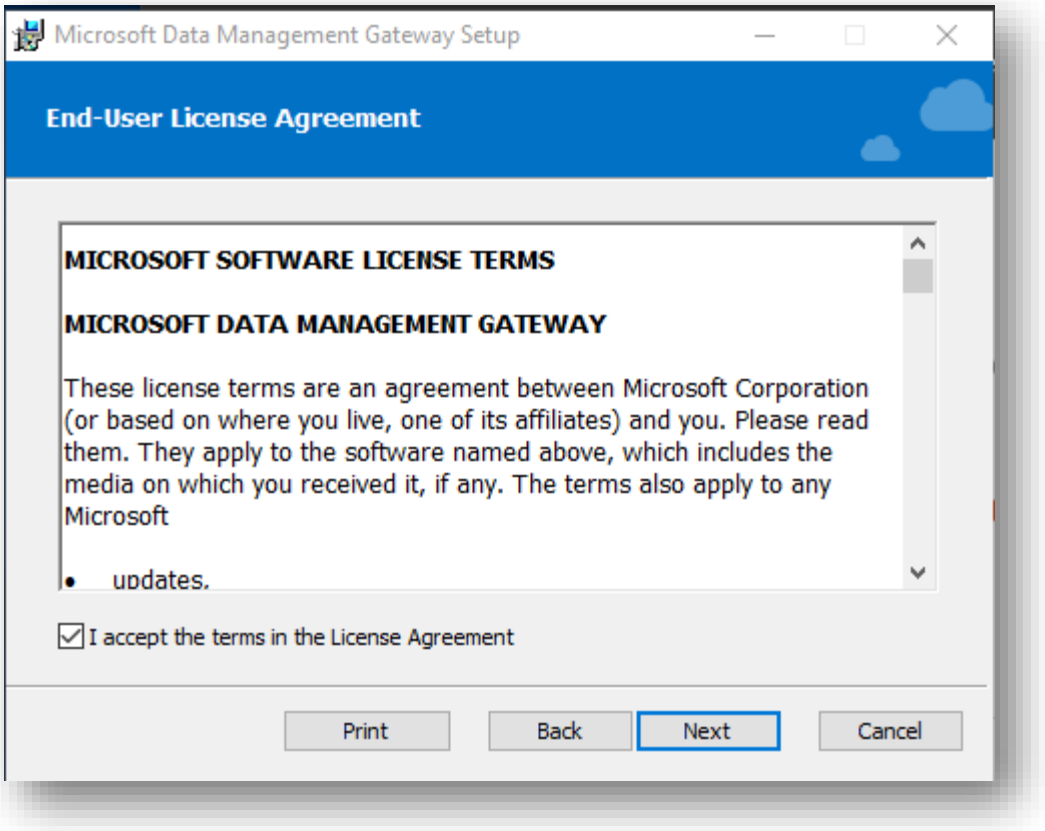
Conclusion

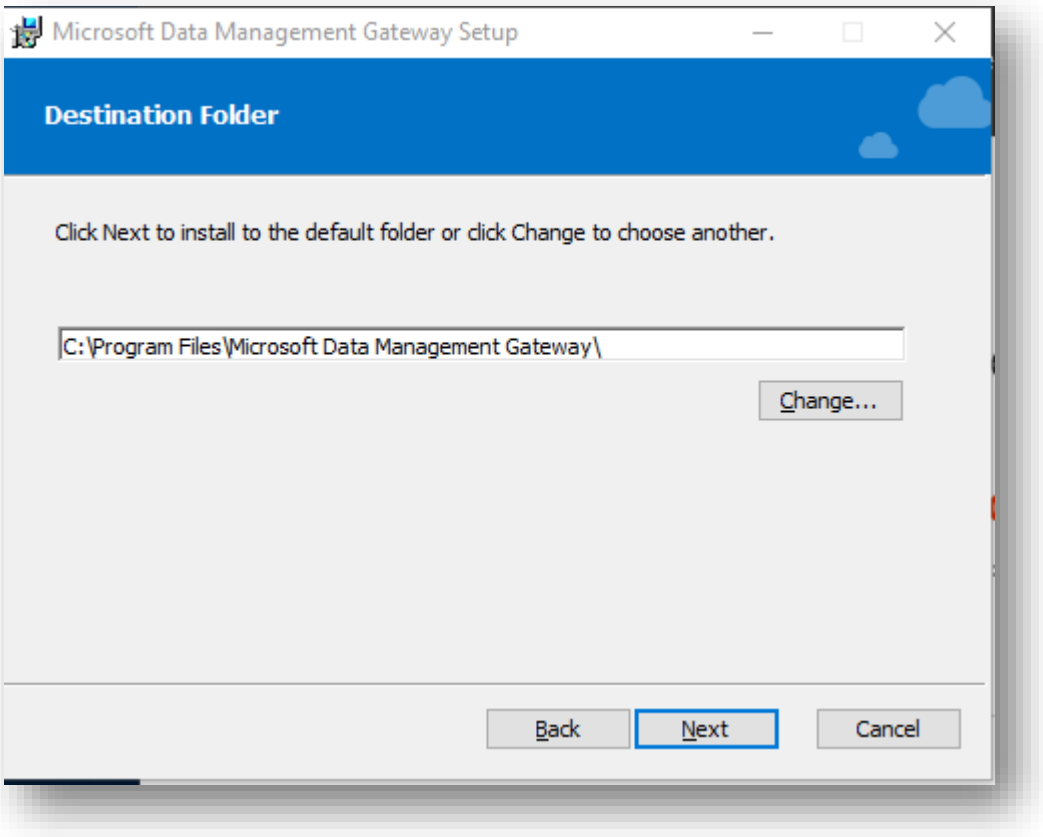
Through this lab, you have learnt how to copy over your files one-time from on-premises location using ADF over to Azure Data Lake Store. You can now expand this learning to try out different types of sources, do copying in a recurring fashion, do different transformations etc. by using Azure Data Factory as the tool of choice.

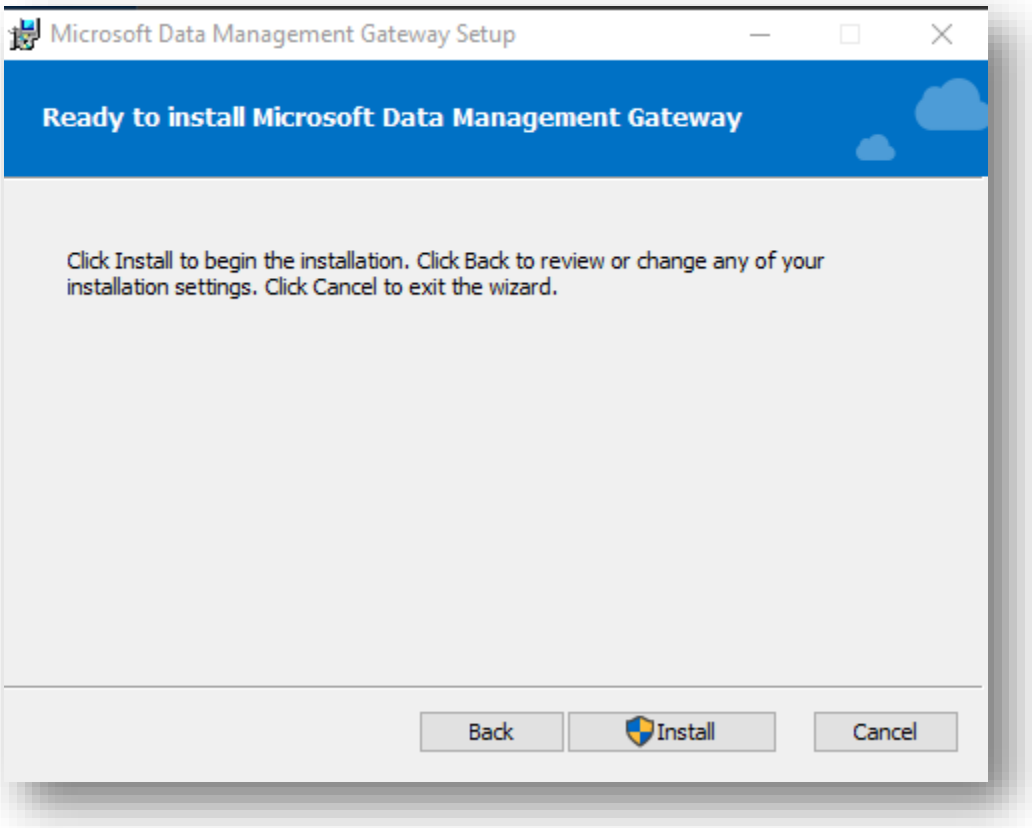
Appendix

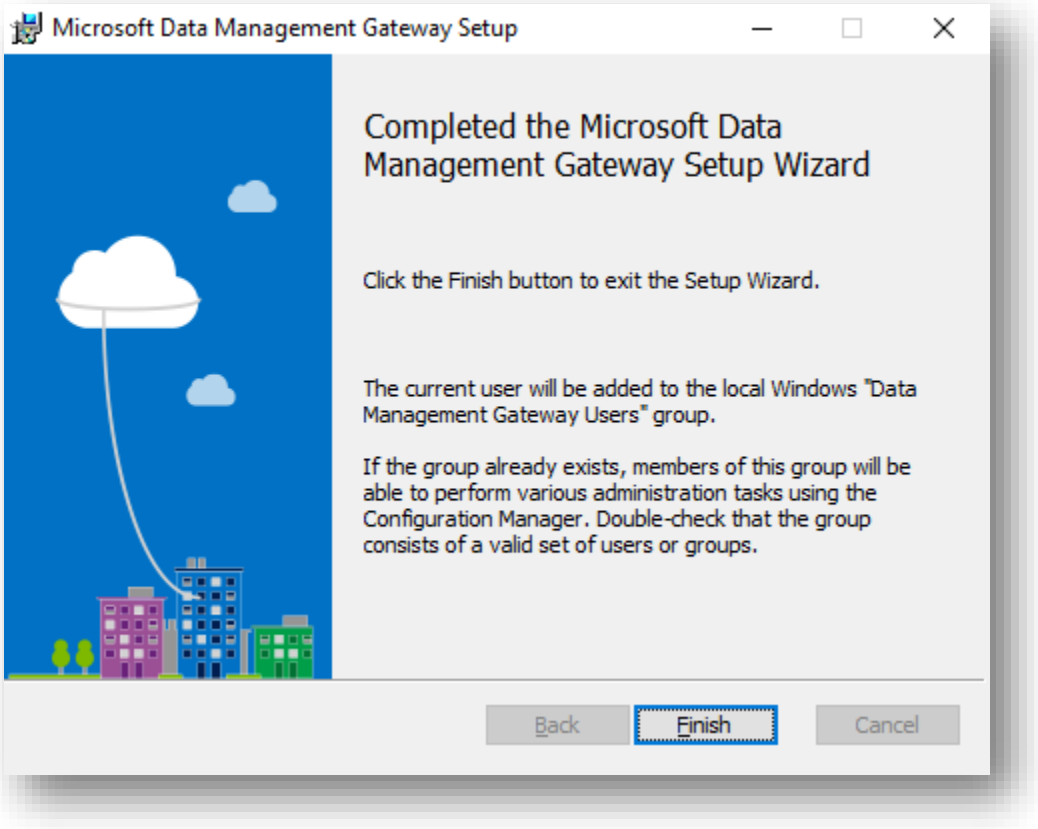
Installing ADF Data Management Gateway manually

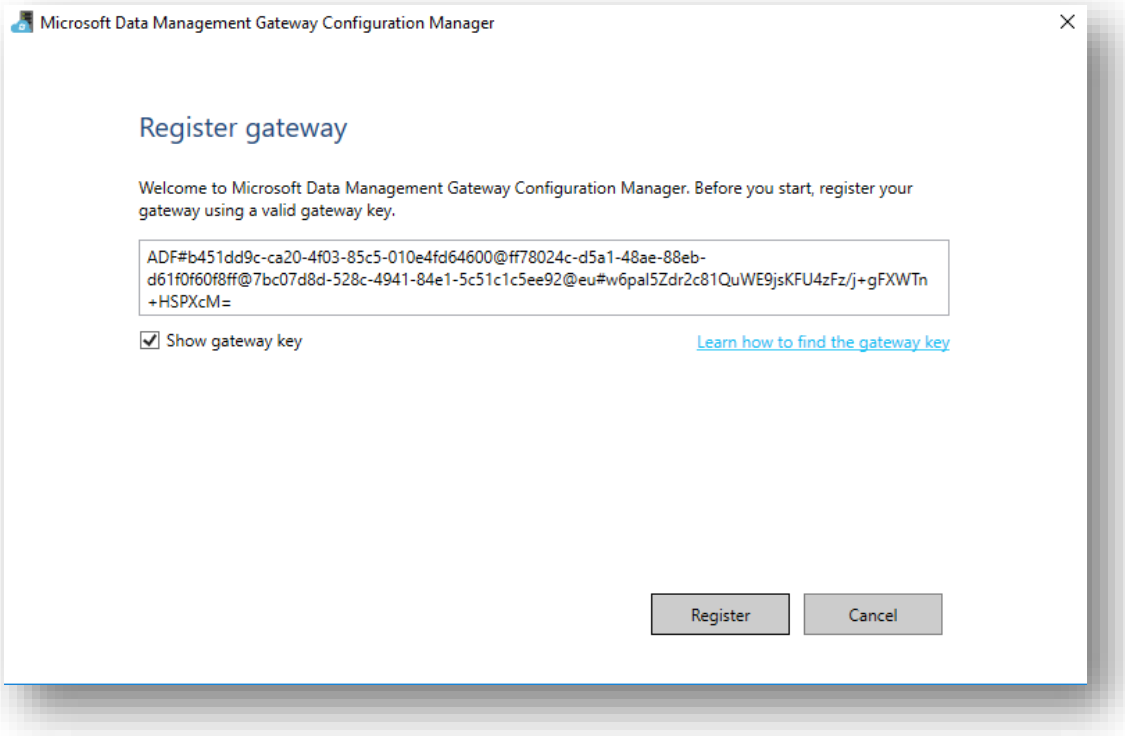



Microsoft Data Management Gateway Setup
End-User License Agreement
MICROSOFT SOFTWARE LICENSE TERMS
MICROSOFT DATA MANAGEMENT GATEWAY
These license terms are an agreement between Microsoft Corporation (or based on where you live, one of its affiliates) and you. Please read them. They apply to the software named above, which includes the media on which you received it, if any. The terms also apply to any Microsoft
updates.
I accept the terms in the License Agreement
Print
Back
Next
Cancel


Microsoft Data Management Gateway Setup
Destination Folder
Click Next to install to the default folder or click Change to choose another.
C:\Program Files\Microsoft Data Management Gateway\
Change...
Back
Next
Cancel


Microsoft Data Management Gateway Setup
Ready to install Microsoft Data Management Gateway
Click Install to begin the installation. Click Back to review or change any of your installation settings. Click Cancel to exit the wizard.
Back
Install
Cancel


Microsoft Data Management Gateway Setup
Completed the Microsoft Data Management Gateway Setup Wizard
Click the Finish button to exit the Setup Wizard.
The current user will be added to the local Windows "Data Management Gateway Users" group.
If the group already exists, members of this group will be able to perform various administration tasks using the Configuration Manager. Double-check that the group consists of a valid set of users or groups.
Back
Finish
Cancel


Microsoft Data Management Gateway Configuration Manager
Register gateway
Welcome to Microsoft Data Management Gateway Configuration Manager. Before you start, register your gateway using a valid gateway key.
ADF#b451dd9c-ca20-4f03-85c5-010e4fd64600@ff78024c-d5a1-48ae-88eb-d61f0f60f8ff@7bc07d8d-528c-4941-84e1-5c51c1c5ee92@eu#w6pal5Zdr2c81QuWE9jsKFU4zFz/j+gFXWTn+HSPXcM=
Show gateway key
Learn how to find the gateway key
Register
Cancel