DATABRICKS HANDS-ON LAB

Microsoft

OVERVIEW

Most companies already have one or more data warehouses. However, extending and maintaining this data warehouse can be difficult. Source systems are changing faster than ever before, and end users want to make deeper analyses.

Therefore, a more flexible architecture is needed which makes it easier to add different types of data.

During this workshop you will experience extend the data warehouse using the Azure Data Services.

The use case during this workshop is about airdelays and preparing the data for Data Scientists on the one hand but also providing it for analysts via the Data Warehouse.

This Lab will guide through the data acquisition and how to create data pipelines with Azure Data Factory and load data into an Azure Blob Storage for further usage with Azure Databricks.

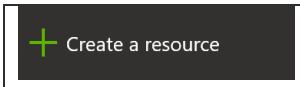
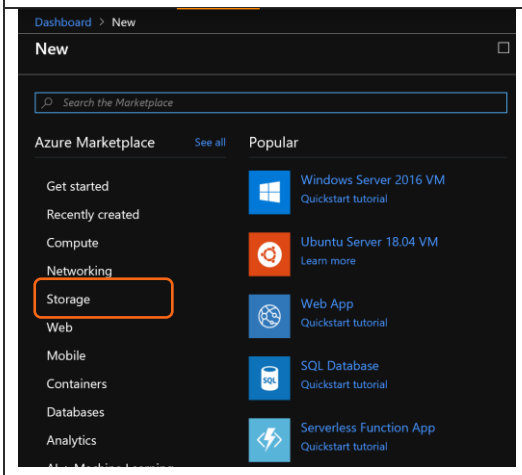
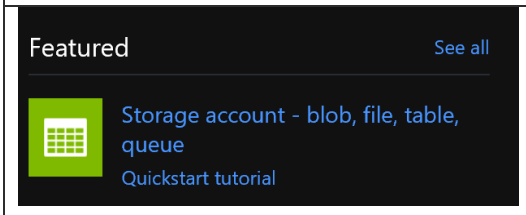
Good luck and enjoy the work!

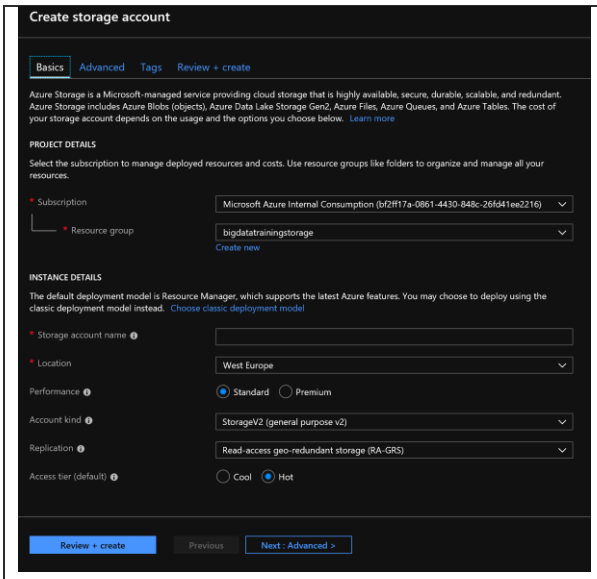
LAB 1: CREATE THE DATA FACTORY AND LOAD THE DATA TO AN AZURE BLOB STORAGE

OVERVIEW

This first lab for today will walk you through the creation of a Azure Blob Storage and the use of Azure Data Factory to fetch files from a folder on Azure, that we then will use as the data in all the subsequent labs.

CREATE AN AZURE BLOB STORAGE

	On the Azure Portal pleas search for the 'Create a resource' button.
	Within the opening blade you will find the entry for 'Storage'
	Please go for 'Storage account – blob, …'



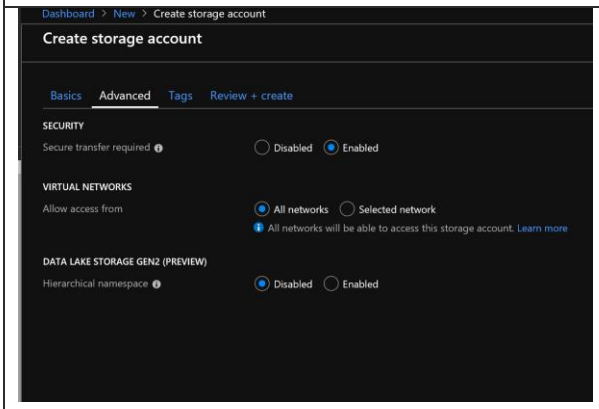
In the first drop-down box please search for your subscription.

In the second box please create a new resource group. This group can be used throughout the whole training. Afterwards you can delete all created artifacts by just deleting this resource group.

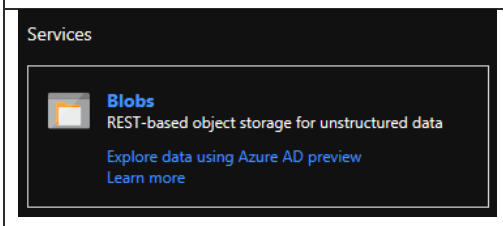
Please provide a 'Storage account name'. This name must be unique over all the storage accounts in Azure and can only contain lower letters and numbers.

When choosing the location for the service please take care, that all services created today are in the same region to avoid additional data egress cost.

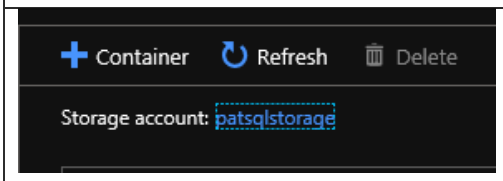
Leave the rest of the properties as is and hit 'Next: Advanced >'



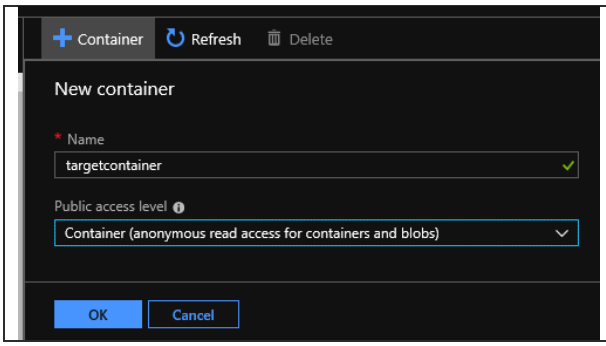
On this blade, leave all entries on their default values and hit 'Review + create' ➔ the storage account will be created for you.



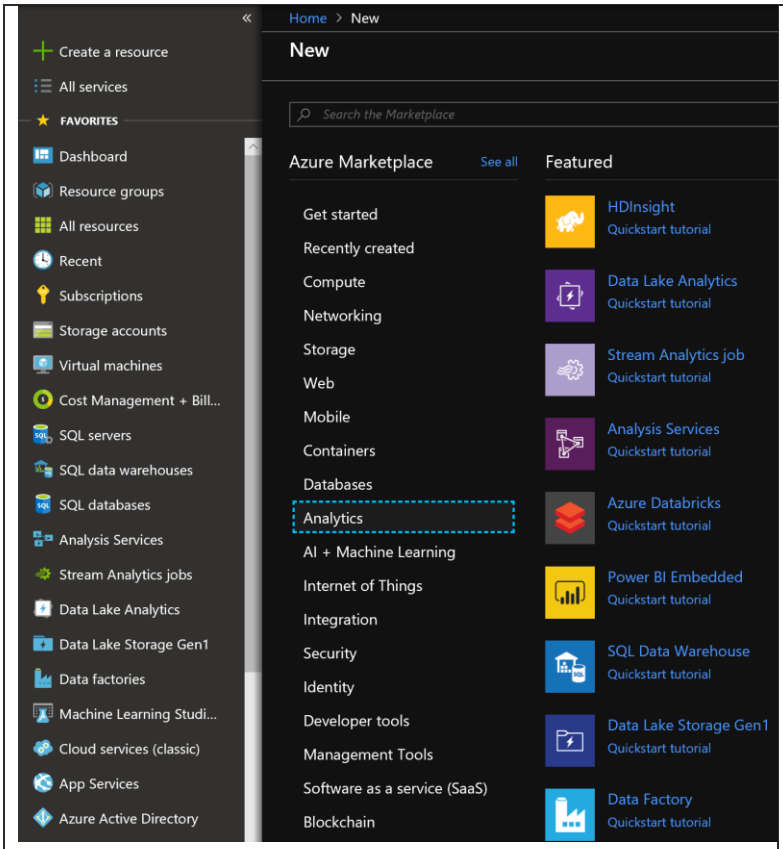
After the storage account is created, please open it and select 'Blobs'



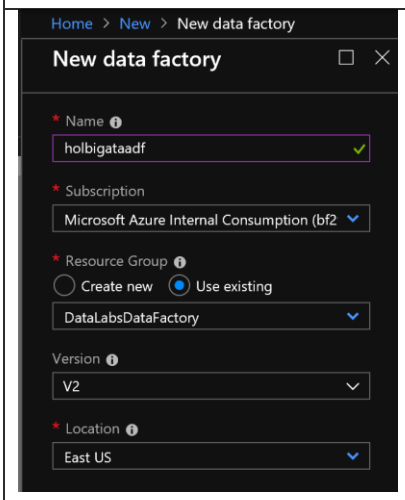
Please click on '+ Container' to create a new container, to store the incoming data.

	Name it and then adjust the 'Public access level' like in the screenshot and hit 'OK'. This will be the target folder for the next step.

CREATE AN AZURE DATA FACTORY AND LOAD FILES FROM A REMOTE STORAGE TO OUR BLOB STORAGE



Go to your Azure portal and click +, choose Analytics then choose Data Factory

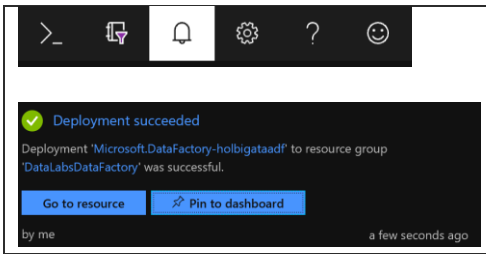
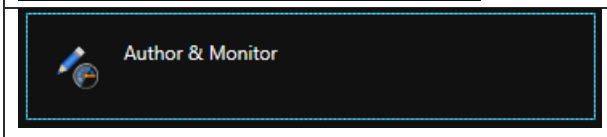
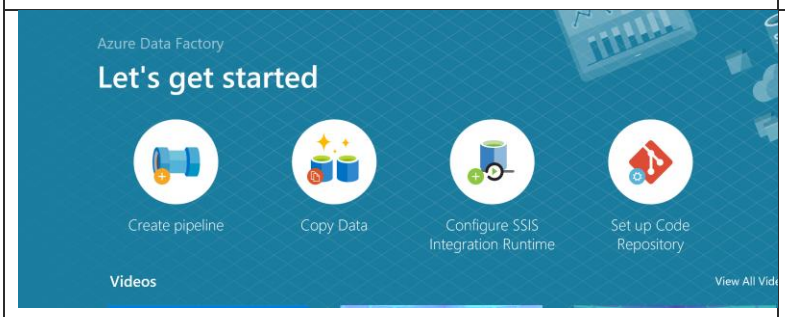
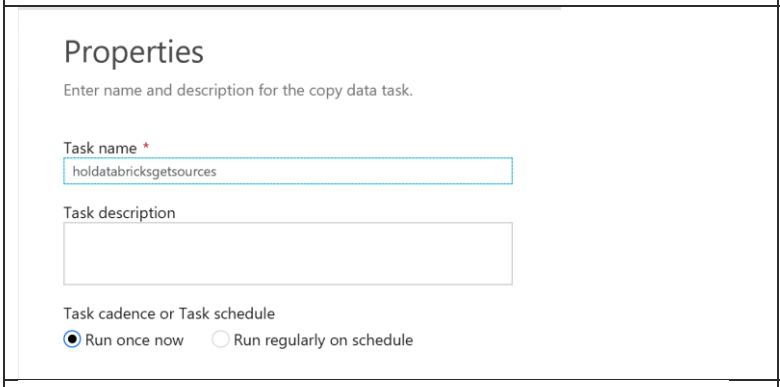
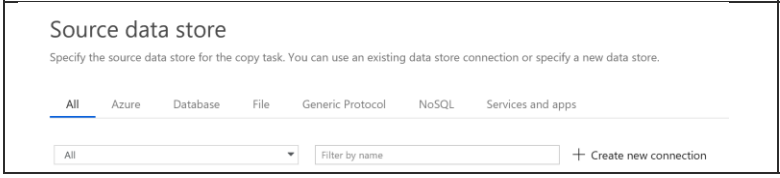


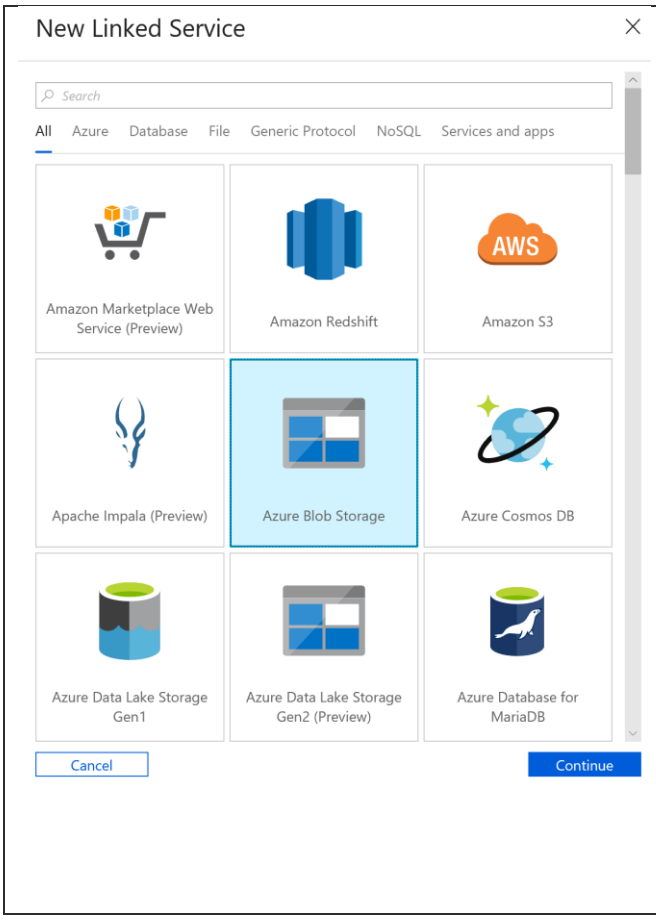
Enter a unique name.

Choose the resource group you have created in the last sequence for your storage service

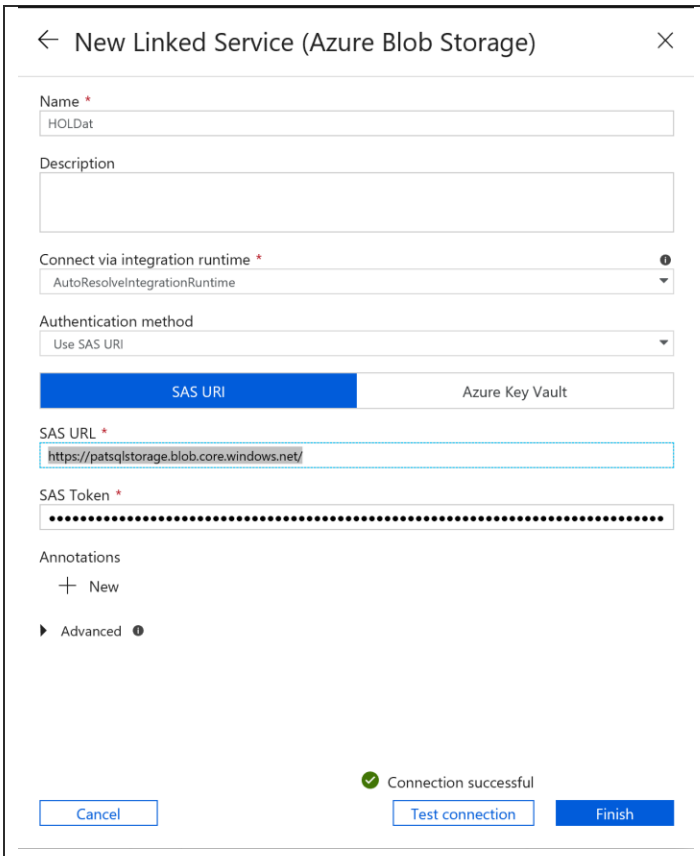
You want to use Version 'V2' for this lab.

And don't forget to put the Data Factory into the region, that you chose for the storage account above. (it will definitely differ from the one in the picture 😉)

	Hit 'create' and wait for the service to be displayed. After creation, the Bell Sign will show new messages. Click on it, choose the Deployment-Success message for your Data Factory and click 'Pin to dashboard'.
	Now open your Data Factory: search for 'Author and Monitor'
	and in the new tab (takes a short amount of time) start the editor: 'Copy Data'.
	On the first screen name your data copy pipeline like 'holdatabricksgetsourcedata' or just leave the default name and click 'Next'
	Here we will create a new data store, that will be used as a connection within Data Factory Pipelines. In this case it'll be the source data for your pipeline.



Click on 'Azure Blob Storage' and hit 'Continue'



On this dialog please name your connection.

Please adjust the 'Authentication method' to 'Use SAS URI'.

For the SAS URL please use:

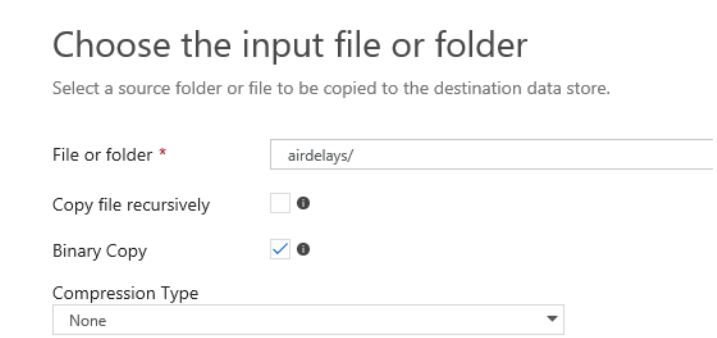
https://patsqlstorage.blob.core.windows.net/

and for the SAS Token please use:

VCXx0TmAO3iMVAY+IOHtV0YXvpsB8WFni1/QcqEYjCAAUComxLDTyiQ65t3A961FBoypsbq9AcRVizsSUaFGog==

You don't need to add other information and can klick on 'Test connection' to check connectivity. If the connection turns green, you may hit 'Finish'.

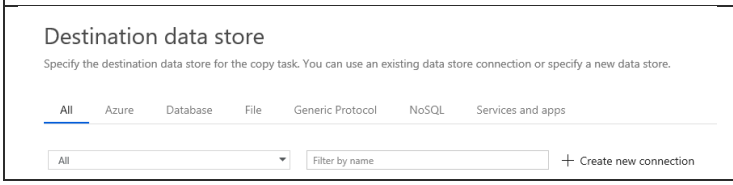
With the newly created source marked, please hit 'Next'.



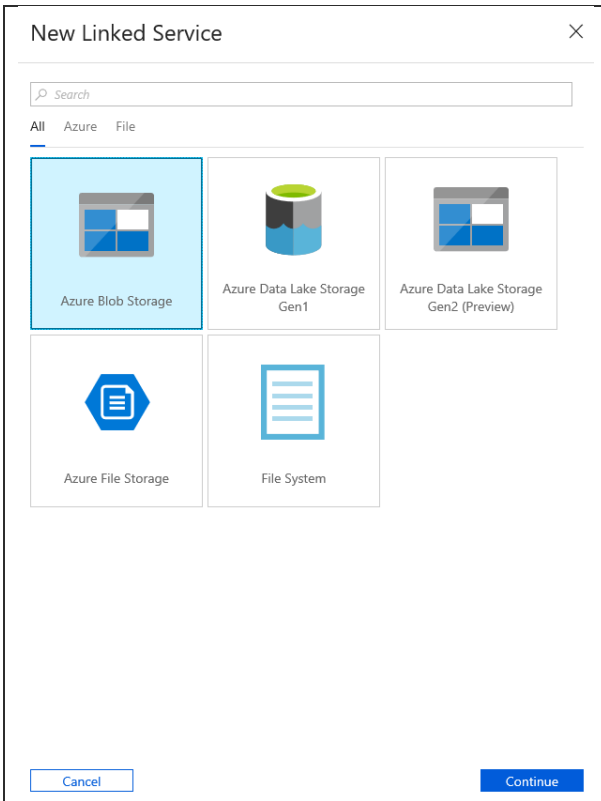
On the following screen please select the source folder ➔ click 'Browse', select the folder and click 'Choose'.

Then please check the 'Binary Copy' property.

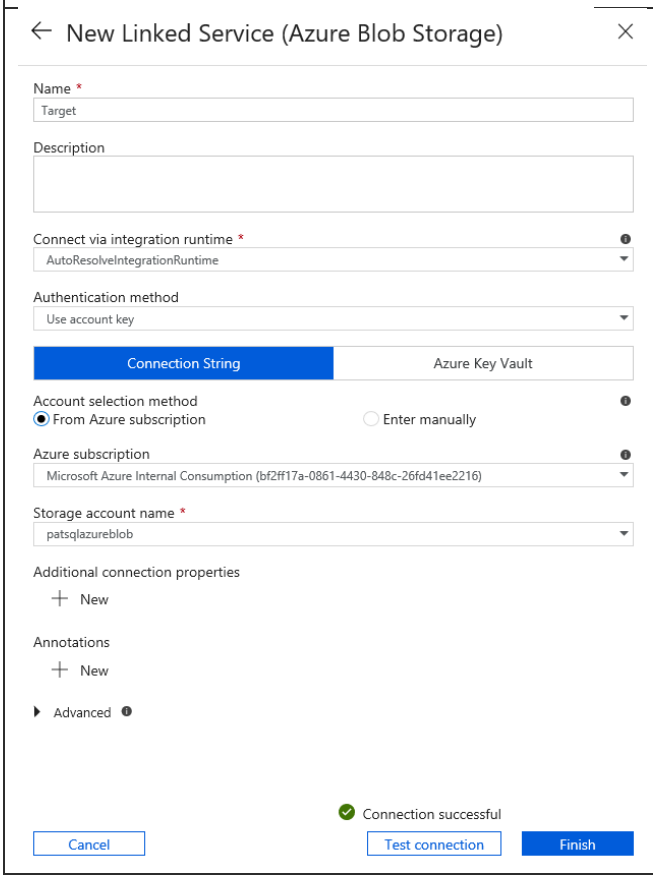
Hit 'Next'



On the 'Destination data store' screen please again create another connection, this time to the storage account you have created above.



Please create a new Blob Storage connection.



Name the target connection.

Please leave 'Use account key' in the 'Authentication method' box.

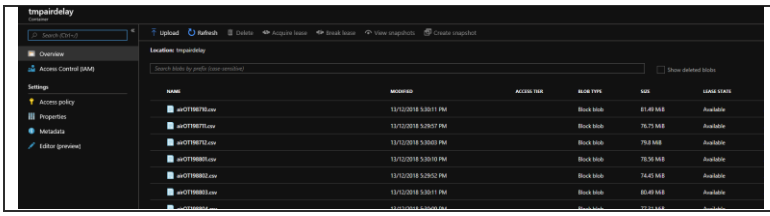
As the 'Account selection method' please also leave 'From Azure subscription'.

In the 'Azure subscription' selector please pick your subscription.

In the 'Storage account name' box select the account, you have created above.

If the connection tests successful, please hit 'Finish' and with the newly created source selected click 'Continue' and proceed to the folder picker dialog.

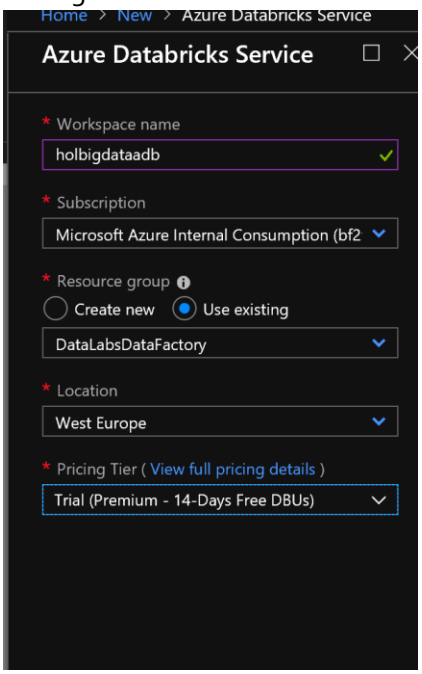
Choose the output file or folder Specify a folder that will contain output files or a specific output file in the destination data store. Folder path * tmpairdelay File name Filenames are defined by source Compression Type None Copy behavior Preserve hierarchy	Choose the folder created above and hit 'Next'
Settings More options for data movement Performance settings Enable Staging Advanced settings	This dialog can be left like it is. (if you would like to stage data, this would be the place to configure a folder for staging files before they're loaded to another Azure Data Service).
Summary You are running pipeline to copy data from Azure Blob Storage to Azure Blob Storage. Azure Blob Storage → Azure Blob Storage Properties Edit Task name CopyPipeline_g1v Task description Source Edit Connection name AzureBlobStorage1 Dataset name SourceDataset_g1v File name Directory path airdelays Destination Edit Connection name Target Dataset name DestinationDataset_g1v File name Directory path tmpairdelay Copy settings Edit Timeout 7.00:00:00 Previous Next	On the 'Summary' blade you will have the last chance to check and change settings.
Azure Blob Storage → Azure Blob Storage Deployment complete Creating Datasets Creating Pipelines Running Pipelines Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close the copy wizard. Edit Pipeline Monitor	After hitting 'Next' the pipeline will be deployed to the Data Factory. You will have the chance to monitor the pipeline execution.

	Now you can proceed to the target folder and check the data, that was loaded. It is now available for the processing with Databricks.

CREATE YOUR DATABRICKS CLUSTER

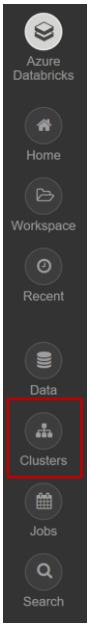
In the next step you will create an Azure Databricks cluster for processing the data that you have loaded.

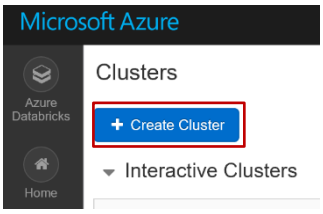
1. Please go to the portal, click '+ Create a resource' and on the Analytics-Tab search for 'Azure Databricks'
2. On the following dialogue please enter a unique name for your Databricks Cluster, the resource group that you have created in the first lab and select the region 'Western Europe' and the Pricing Tier:



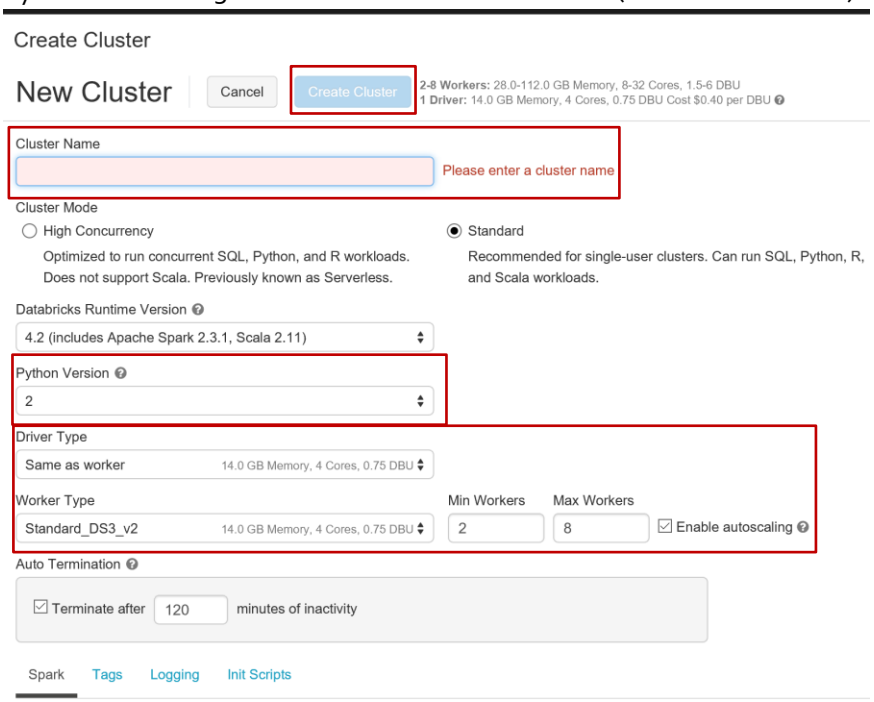
3. After the Databricks workspace is created and shows up on the dashboard please select it and click on 'Launch Workspace'. It will show up in a different tab:

4. As the workspace launches we can start and create the first Databricks cluster. Please click 'Clusters' and then on the next screen 'Create Cluster'

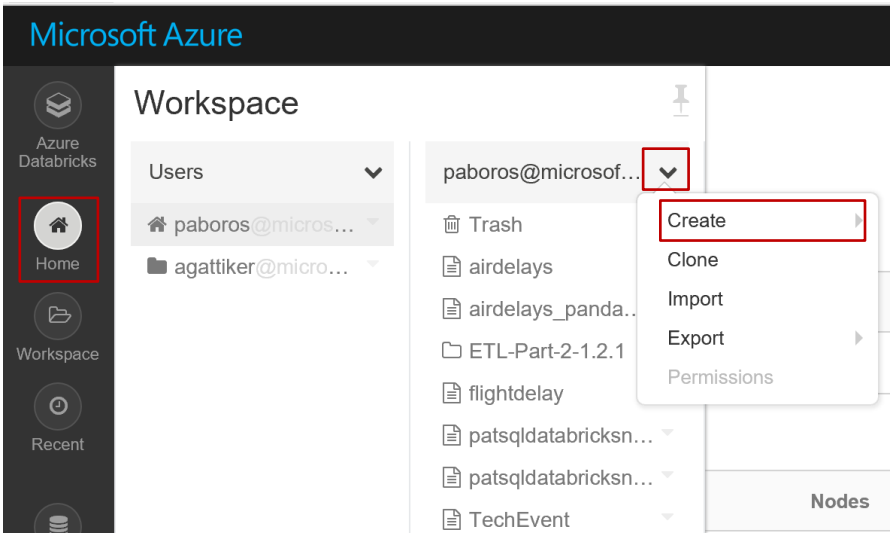




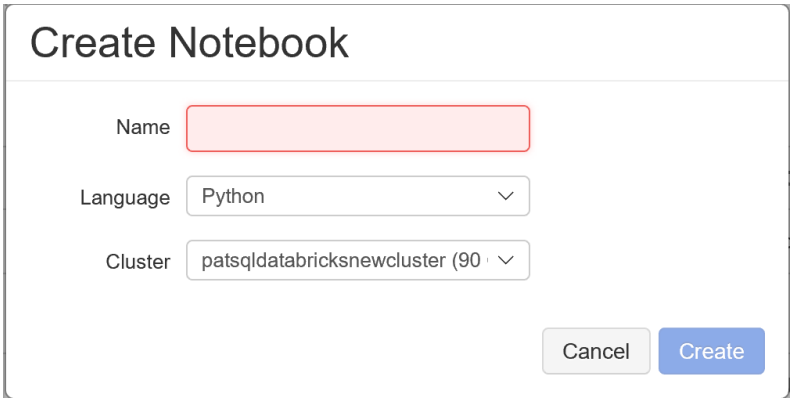
5. On the following screen you will be able to configure the planned cluster as you need it. Please enter a Cluster Name, select the Databricks Runtime Version (leave the default), select the Python Version = 3 and the size of the Cluster-VMs (Driver and Worker) and hit 'Create Cluster'



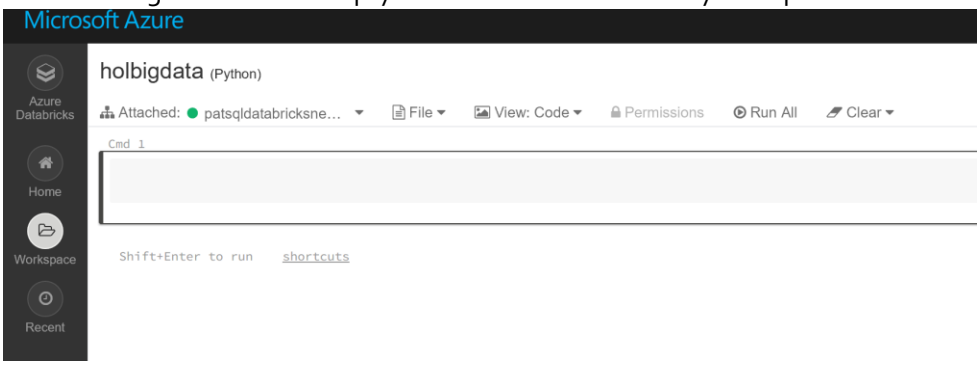
6. To use the Cluster we need to create a notebook now, where we can enter code and run it against it. Please click on 'Home' and on the Users tab on the arrow. Then select 'Create'



7. On the next cartridge select 'Notebook'. Name the new notebook. The rest can be left defaulted



8. After clicking 'Create' the empty notebook is available for your input.



TERMS OF USE

By using this Hands-on Lab, you agree to the following terms:

The technology/functionality described in this Hands-on Lab is provided by Microsoft Corporation in a “sandbox” testing environment for purposes of obtaining your feedback and to provide you with a learning experience. You may only use the Hands-on Lab to evaluate such technology features and functionality and provide feedback to Microsoft. You may not use it for any other purpose. You may not modify, copy, distribute, transmit, display, perform, reproduce, publish, license, create derivative works from, transfer, or sell this Hands-on Lab or any portion thereof.

COPYING OR REPRODUCTION OF THE HANDS-ON LAB (OR ANY PORTION OF IT) TO ANY OTHER SERVER OR LOCATION FOR FURTHER REPRODUCTION OR REDISTRIBUTION IS EXPRESSLY PROHIBITED.

THIS HANDS-ON LAB PROVIDES CERTAIN SOFTWARE TECHNOLOGY/PRODUCT FEATURES AND FUNCTIONALITY, INCLUDING POTENTIAL NEW FEATURES AND CONCEPTS, IN A SIMULATED ENVIRONMENT WITHOUT COMPLEX SET-UP OR INSTALLATION FOR THE PURPOSE DESCRIBED ABOVE. THE TECHNOLOGY/CONCEPTS REPRESENTED IN THIS HANDS-ON LAB MAY NOT REPRESENT FULL FEATURE FUNCTIONALITY AND MAY NOT WORK THE WAY A FINAL VERSION MAY WORK. WE ALSO MAY NOT RELEASE A FINAL VERSION OF SUCH FEATURES OR CONCEPTS. YOUR EXPERIENCE WITH USING SUCH FEATURES AND FUNCITONALITY IN A PHYSICAL ENVIRONMENT MAY ALSO BE DIFFERENT.

FEEDBACK. If you give feedback about the technology features, functionality and/or concepts described in this Hands-on Lab to Microsoft, you give to Microsoft, without charge, the right to use, share and commercialize your feedback in any way and for any purpose. You also give to third parties, without charge, any patent rights needed for their products, technologies and services to use or interface with any specific parts of a Microsoft software or service that includes the feedback. You will not give feedback that is subject to a license that requires Microsoft to license its software or documentation to third parties because we include your feedback in them. These rights survive this agreement.

MICROSOFT CORPORATION HEREBY DISCLAIMS ALL WARRANTIES AND CONDITIONS WITH REGARD TO THE HANDS-ON LAB, INCLUDING ALL WARRANTIES AND CONDITIONS OF MERCHANTABILITY, WHETHER EXPRESS, IMPLIED OR STATUTORY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. MICROSOFT DOES NOT MAKE ANY ASSURANCES OR REPRESENTATIONS WITH REGARD TO THE ACCURACY OF THE RESULTS, OUTPUT THAT DERIVES FROM USE OF THE VIRTUAL LAB, OR SUITABILITY OF THE INFORMATION CONTAINED IN THE VIRTUAL LAB FOR ANY PURPOSE.

DISCLAIMER

This lab contains only a portion of the features and enhancements in Microsoft Azure Data Factory, Azure SQL Data Warehouse, Azure DataBrics and Azure Data Lake Storage. Some of the features might change in future releases of the product.