

# BIG DATA HANDS-ON LAB

Azure Data Services:

Data Factory

Data Lake Store

SQL Data Warehouse

Databricks

Overview.....	3
Summary .....	3
Prerequisites .....	3
Lab 1: Create the data Warehouse and load a dimension table .....	4
Overview.....	4
Create a SQL Datawarehouse .....	4
Scalability of SQL Data Warehouse in the Azure Portal.....	8
Analyze customer data with data engineering tools.....	10
Create an Azure Data Factory and Load a file from the Internet into our SQL Data Warehouse.....	15
Lab 2: Creating and Loading a Data Lake Gen 2 .....	24
Overview.....	24
Pre-load your data to Data Lake.....	24
Lab 3 Connect with Storage Explorer to see the storage accounts. ....	30
Lab 4.....	34
Terms of use .....	34

## OVERVIEW

### SUMMARY

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 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.

Lab 1 will guide through the data acquisition and how to create data pipelines with Azure Data Factory and load data into Azure SQL Datawarehouse

Lab 2 guides through the creation and usage of a Databricks Cluster. You will use Python and SQL to analyze and massage the data and provide it for further usage with other services.

Lab 3 will then examine the possibilities with Azure SQL Data Warehouse and provides some insights into the world of MPP- (massive parallel processing) databases. You will get data with Polybase from your Data Lake and join a dimension table that lives in the database.

Good luck and enjoy the labs!

### PREREQUISITES

SQL Server Management Studio: <https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms>

Azure Storage Explorer: <https://azure.microsoft.com/en-us/features/storage-explorer/>

This Link will be needed later on in the lab. (Don't click, it won't take you anywhere 😊)

<https://patsqlstorage.blob.core.windows.net/?sv=2017-11-09&ss=bfqt&srt=sco&sp=rl&se=2018-11-16T20:17:43Z&st=2018-11-16T12:17:43Z&spr=https&sig=ejhXpG8RJ6KF5Tu2jxT7y%2FXT9nZfGL%2BAVAqTIA4Uzdw%3D>

## LAB 1: CREATE THE DATA WAREHOUSE AND LOAD A DIMENSION TABLE

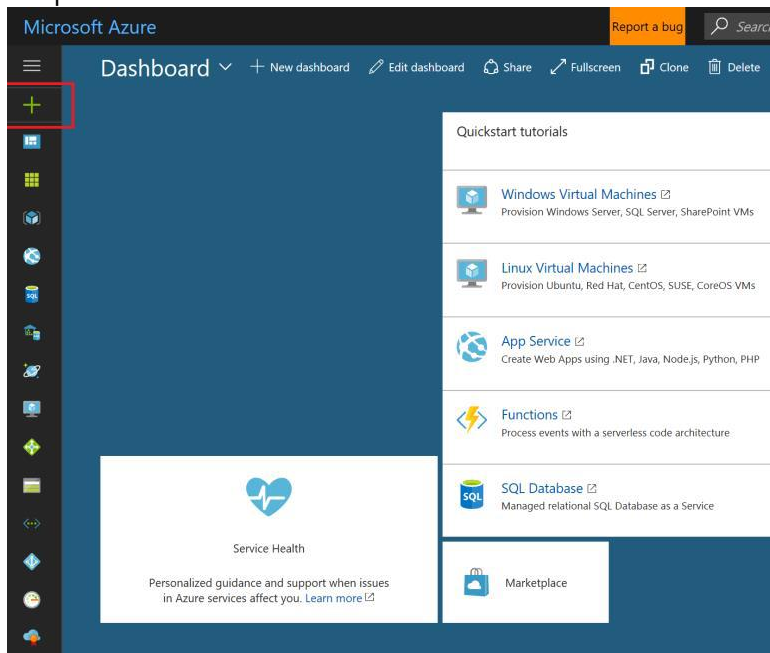
### OVERVIEW

This first lab for today will walk you through the creation of a SQL Data Warehouse and the use of Azure Data Factory to fetch a file from the web, that we then will use as a Dimension table in the subsequent labs.

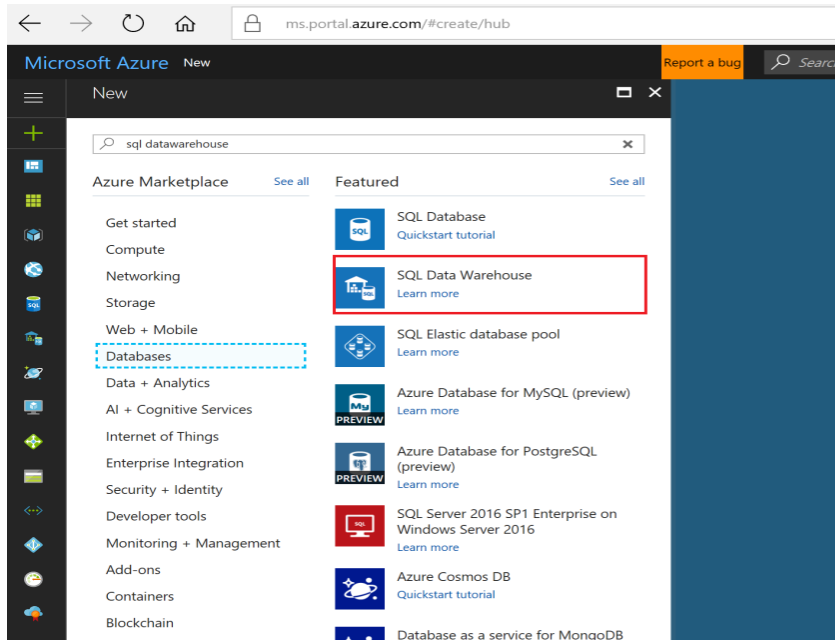
### CREATE A SQL DATAWAREHOUSE

In this step we will walk through the creation of a SQL Data warehouse.

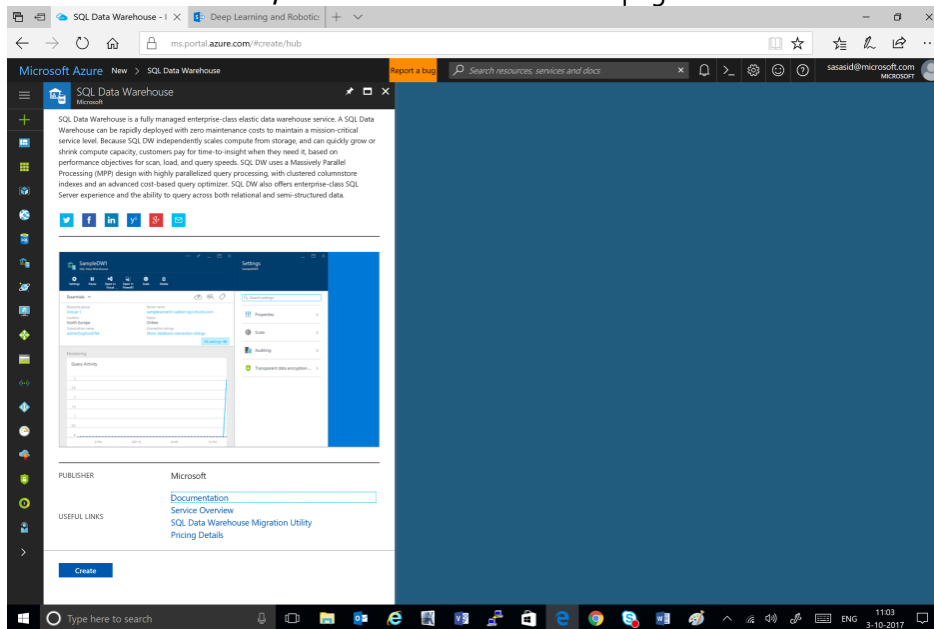
1. Open a browser and go to the Azure Portal: <http://portal.azure.com>. Please login with your user.
2. Next step is to create a new SQL Data warehouse. Click on the "+" sign which you see on your left pane of the window.



3. Select databases and then click on SQL Data Warehouse.



4. Click on create button, this will launch the creation page for SQL Data Warehouse



5. Once you are on this page enter the name of the SQL Data Warehouse you want to create.

Microsoft Azure New > SQL Data Warehouse > SQL Data Warehouse

SQL Data Warehouse

New

Dashboard

All resources

Resource groups

App Services

SQL databases

SQL data warehouses

Azure Cosmos DB

\* Database name  
DataLab ✓

\* Subscription  
Microsoft Azure Internal Consumption ✓

\* Resource group ⓘ  
☐ Create new ☒ Use existing  
DataLab ✓

\* Select source ⓘ  
Sample ✓

Select sample ⓘ  
AdventureWorksDW ✓

6. In the subscription tab click on the drop down and select the subscription which is listed.
7. For Resource Group click on the "create new" link under the Combo Box and provide a name like "datalabrg".
8. Select Sample in the select source option
9. In the next selection, click on the server option. This will prompt you to create a new SQL Server. Under the new Server tab enter the SQL Server name, password and location as West Europe and Click on Select. Please take a note of the user name and password since we will be using this later during the lab for logging into the data warehouse

Microsoft Azure New > SQL Data Warehouse > SQL Data Warehouse > Server > New server

SQL Data Warehouse

New

Dashboard

All resources

Resource groups

App Services

SQL databases

SQL data warehouses

Azure Cosmos DB

Virtual machines

Load balancers

\* Database name  
DataLab ✓

\* Subscription  
Microsoft Azure Internal Consumption ✓

\* Resource group ⓘ  
☐ Create new ☒ Use existing  
DataLab ✓

\* Select source ⓘ  
Sample ✓

Select sample ⓘ  
AdventureWorksDW ✓

\* Server  
Configure required settings >

Server

Create a new server

No servers found

New server

\* Server name  
datalab ✓  
.database.windows.net

\* Server admin login  
datalab ✓

\* Password  
\*\*\*\*\* ✓

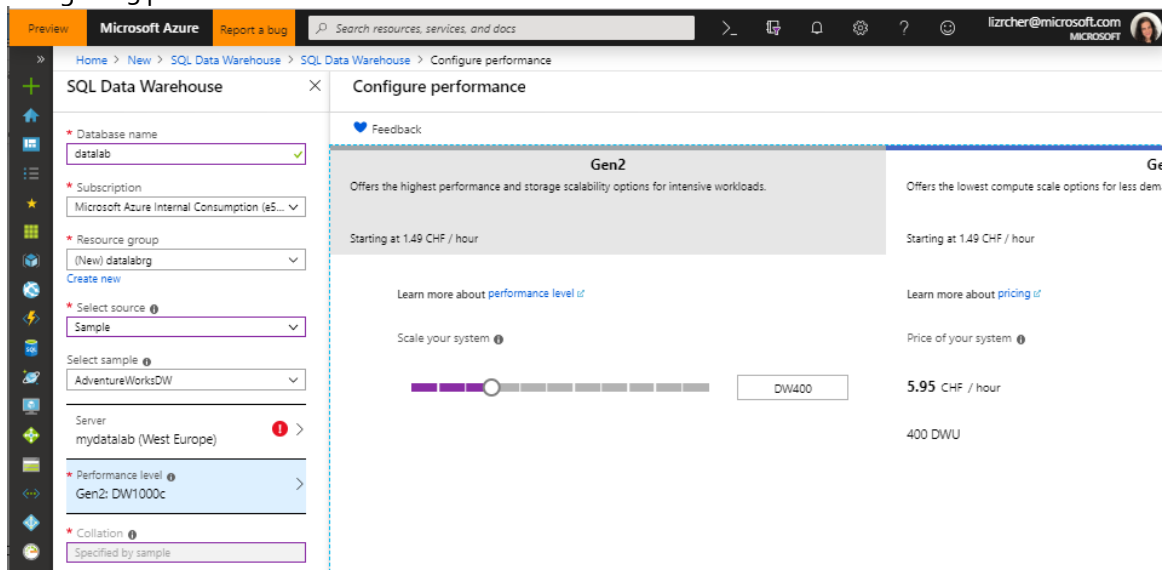
\* Confirm password  
\*\*\*\*\* ✓

\* Location  
West Europe ✓

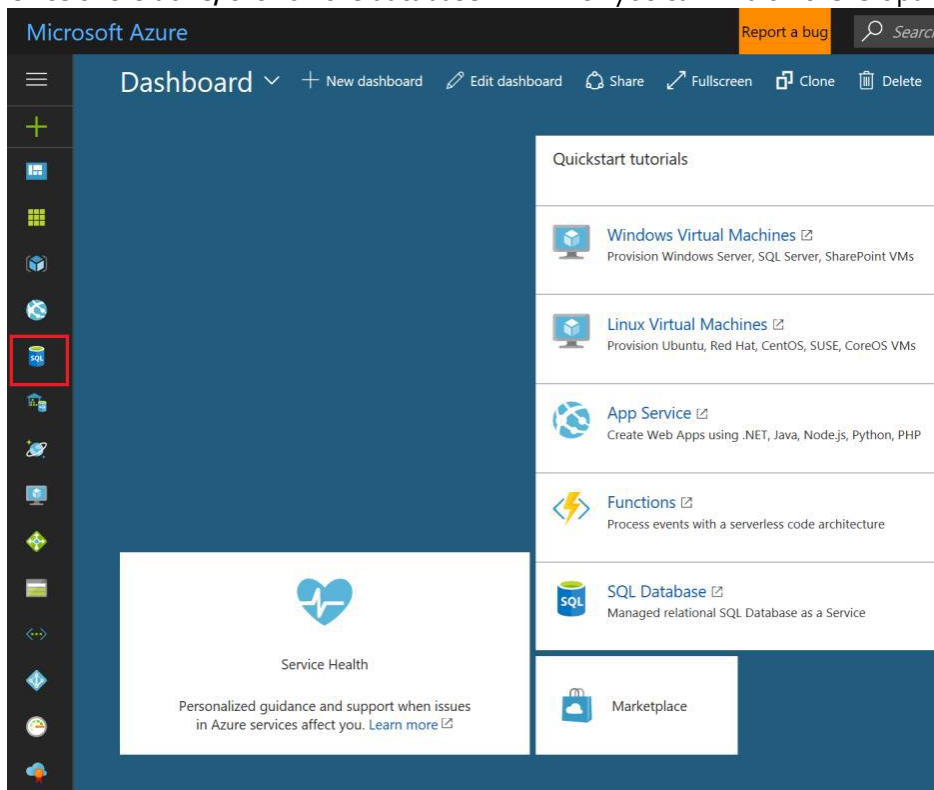
☒ Allow azure services to access server ⓘ

10. In the performance tier select Gen1 and the default of 400 DWU (Data Warehouse Units) for now. This is a measure of performance capacity for Data Warehouse with CPU, IO and Memory

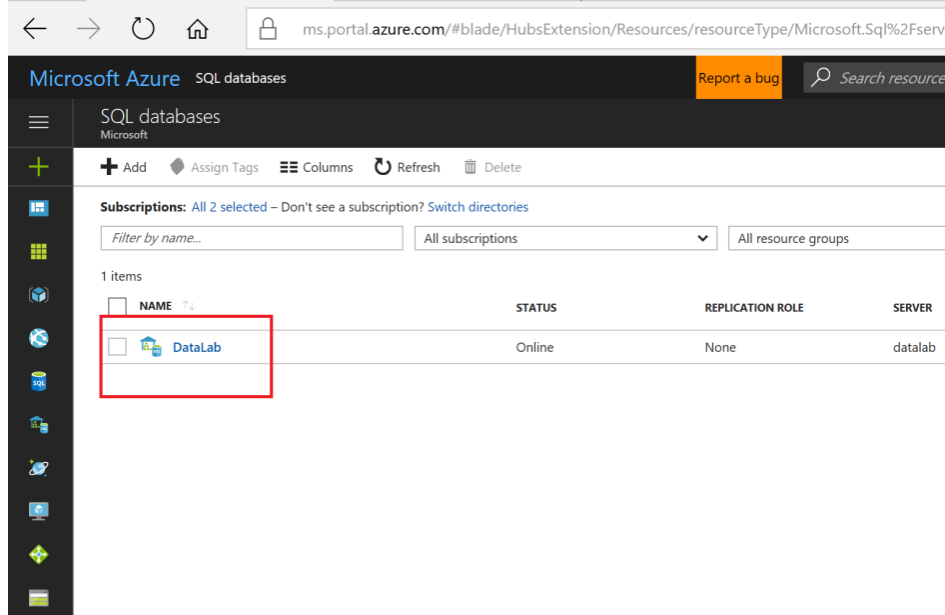
being the 3 parameters which are used to define this unit.



11. On the top right corner of your screen, you will be able to see the notification that the SQL DWH is being deployed. Take your time as this might take some minutes.
12. Once this is done, click on the database link which you can find on the left panel of the portal



13. Click on the Data Warehouse link and this will take you to the overview page



14. You have successfully created a new SQL Data Warehouse.

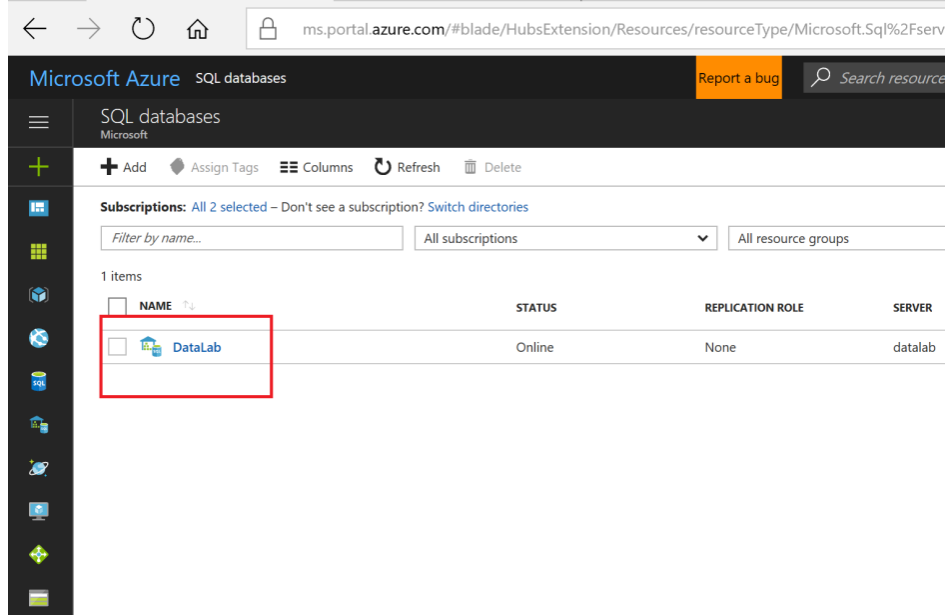
## SCALABILITY OF SQL DATA WAREHOUSE IN THE AZURE PORTAL

In this part of the lab we will connect to a Data Warehouse to get some insight in customer data. In this case it is a relational database designed for Big Data (massive parallel processing technology). On Azure there are also other technologies available for analyzing data, like Hadoop, Python or R.

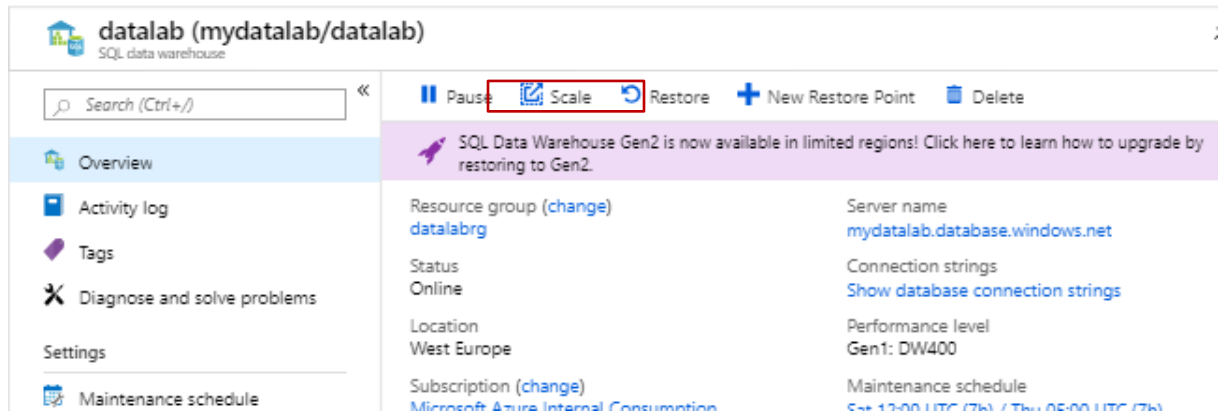
To start, let's explore the scalability features of SQL Data Warehouse, the Azure service is hosting the database.



1. Go to All Resources and when the new pane opens, click on your data warehouse with the name you had provided while creating your Data Warehouse.

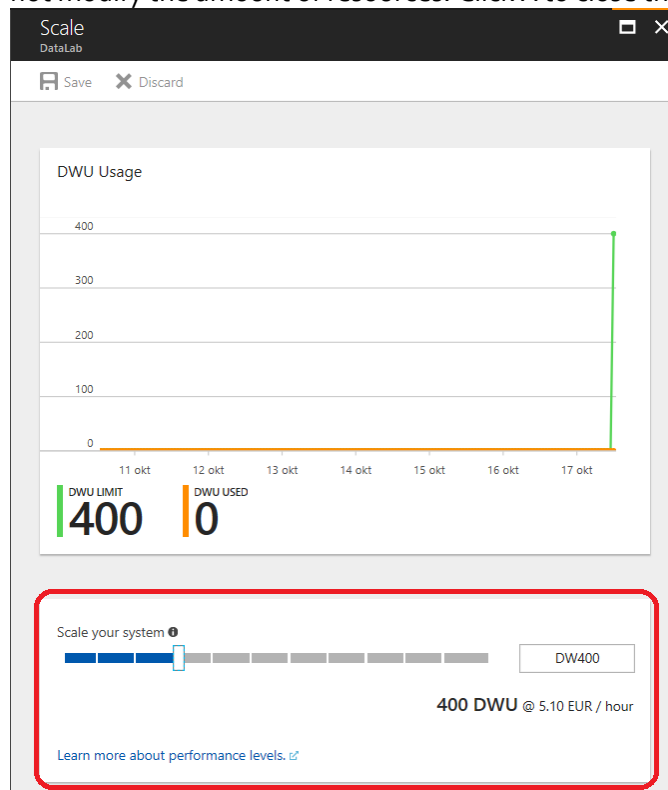


2. Some of the advantages of using Azure SQL Data Warehouse include dynamic scaling and the ability to pause your data warehouse when you are not using it. For example, if you have a period during the day where data is regularly being loaded or processed, you can scale up your data warehouse by increasing the number of DWUs. When the load process finishes, you can scale the data warehouse down by reducing DWUs. Similarly, if there is a time where you will not need compute resources at all, you can pause your data warehouse.
3. Make sure your DWH is running by checking if the status is online. Notice that there is also a button Scale. Click on the Scale button.



4. You will now see a pane where you see the actual usage of the last days. Also, you can increase or decrease the resources of the data warehouse on-demand with a slider. In this lab we'll will

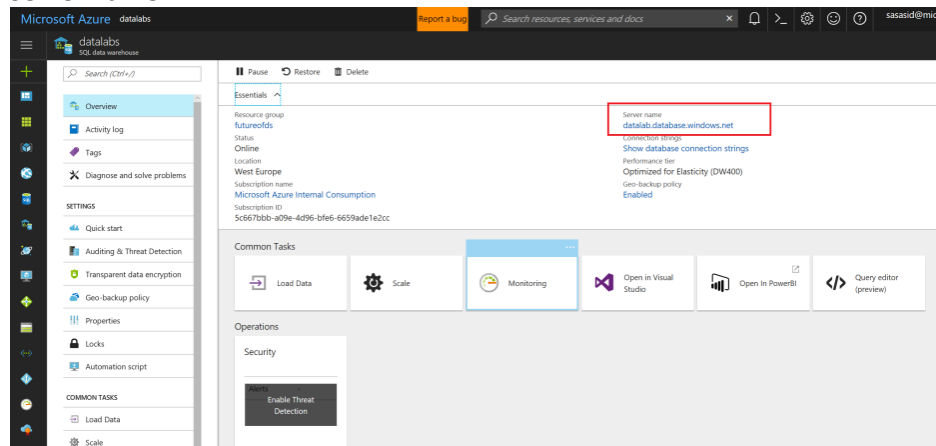
not modify the amount of resources. Click X to close the pane.



## Analyze customer data with data engineering tools

We will execute some queries to the data warehouse by connecting to the Azure SQL Data Warehouse by using SQL Server Management Studio (SSMS). SSMS is usually used by developers and administrators to access SQL Server, Azure SQL Database or Azure SQL Data Warehouse.

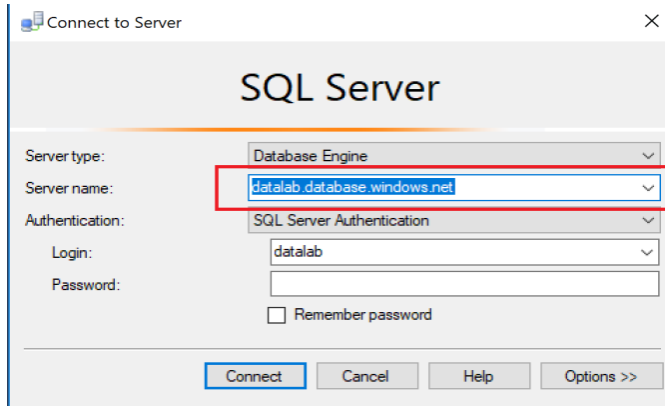
1. On the pane displaying the key information about the Azure SQL Data Warehouse, notice that there is a server name. We'll use this server name to connect to the data warehouse. Copy the server name.



2. Go to your desktop in Windows and open the tool SQL Server Management Studio (SSMS).



3. When opening SSMS, you are asked to connect to a server. Fill in the details as described below and click on Connect.

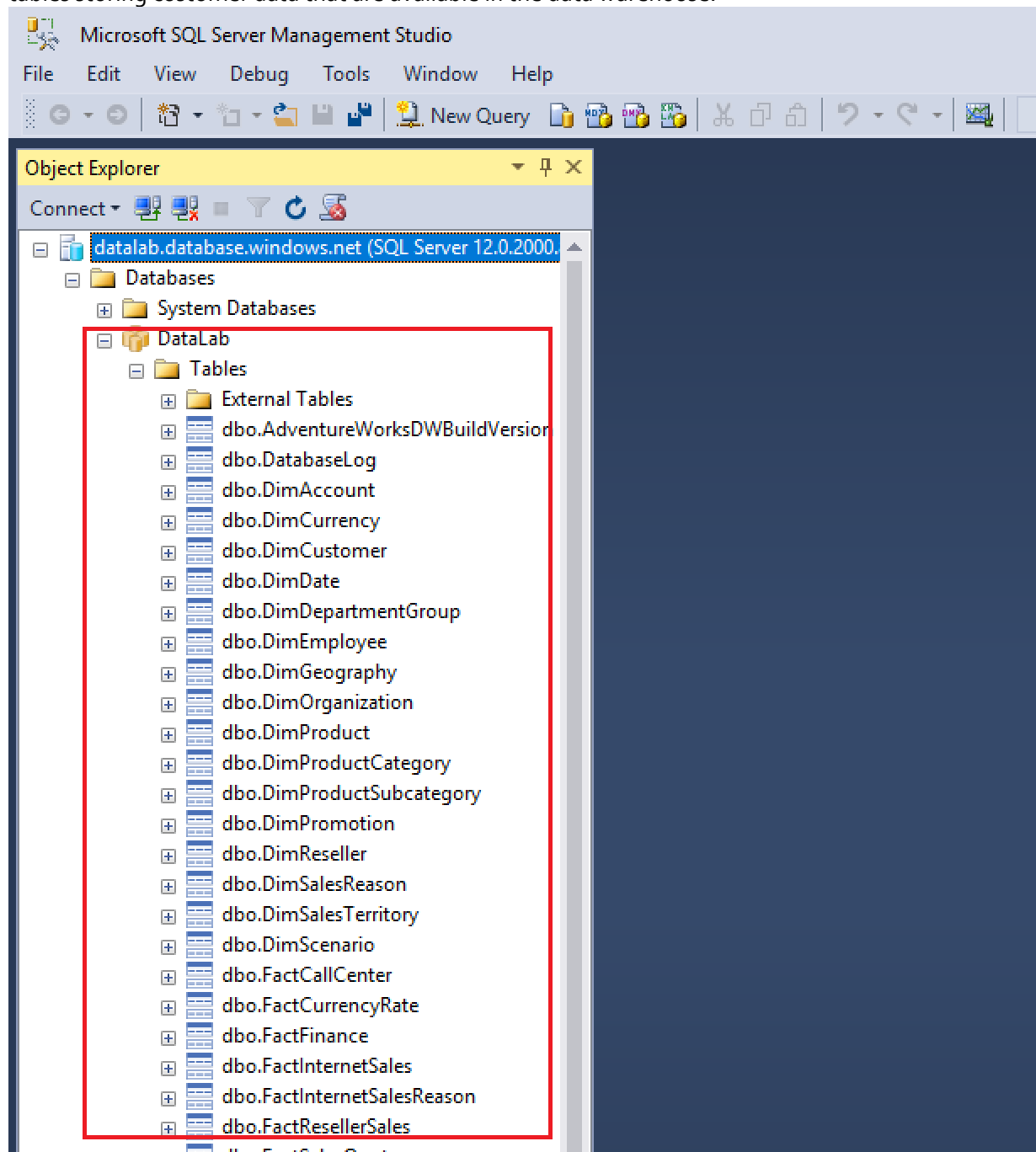


Configuration	Value
Server Type	Database Engine
Server name	[YOURSERVERNAME] (something like datalab.database.windows.net)
Authentication	SQL Server Authentication
Login	[YOURADMINUSER]
Password	[YOURADMINPASSWORD]

If prompted, please on the dialog sign into Azure and then create a new firewall rule by leaving the default of Add my client IP, then click OK. This will create a new rule that adds your IP address, and you should be able to connect to your server.

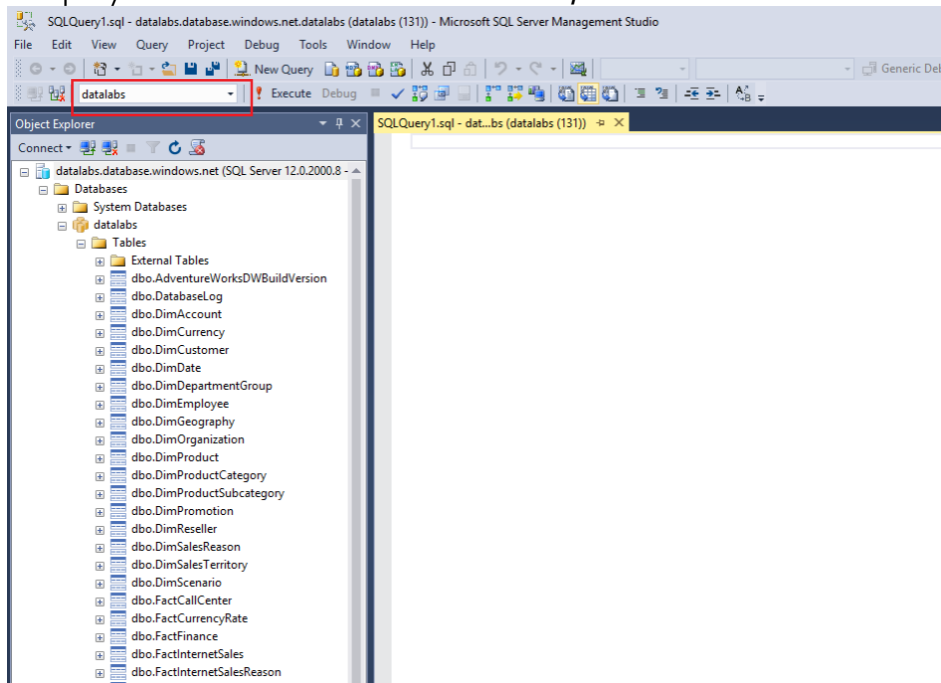
4. When connected, you see in the Object Explorer the databases on the server. When you click on the + the objects in the database will expand. Click on the + of your database and look at the

tables storing customer data that are available in the data warehouse.



5. Let's execute a query against this data warehouse to get some insight in the sales. Click on the New Query button and a new pane will open. Copy and paste the query below in the text window, and make sure that the database name you have used above is select (s. below) to run

the query in instead of “master”. If all is correct, then click on Execute.



```
SELECT [FirstName]
, [LastName]
, [EnglishProductSubcategoryName] [Sub Category]
, SUM([OrderQuantity]) [Order Quantity]
FROM [dbo].[FactInternetSales] sales
INNER JOIN [dbo].[DimCustomer] customer
    ON sales.[CustomerKey] = customer.[CustomerKey]
INNER JOIN [dbo].[DimProduct] AS product
    ON sales.[ProductKey] = product.[ProductKey]
INNER JOIN [dbo].[DimProductSubcategory] subcategory
    ON subcategory.[ProductSubcategoryKey] =
product.[ProductSubcategoryKey]
GROUP BY [FirstName],
[LastName], [EnglishProductSubcategoryName]
ORDER BY SUM([OrderQuantity]) DESC
```

The result of this query shows ordered quantity of products per customer. It will take a couple of seconds to run.

External Tables

- dbo.AdventureWorksDWBldVersion
- dbo.DatabaseLog
- dbo.DimAccount
- dbo.DimCurrency
- dbo.DimCustomer
- dbo.DimDate
- dbo.DimDepartmentGroup
- dbo.DimEmployee
- dbo.DimGeography
- dbo.DimOrganization
- dbo.DimProduct
- dbo.DimProductCategory
- dbo.DimProductSubcategory
- dbo.DimPromotion
- dbo.DimReseller
- dbo.DimSalesReason
- dbo.DimSalesTerritory
- dbo.DimScenario
- dbo.FactCallCenter
- dbo.FactCurrencyRate
- dbo.FactFinance
- dbo.FactInternetSales
- dbo.FactInternetSalesReason
- dbo.FactResellerSales

```

INNER JOIN [dbo].[DimProductSubcategory] subcategory
ON subcategory.[ProductSubcategoryKey] = product.[ProductSubcategoryKey]
GROUP BY [FirstName], [LastName], [EnglishProductSubcategoryName]
ORDER BY SUM([OrderQuantity]) DESC

```

	FirstName	LastName	Sub Category	Order Quantity
1	April	Shan	Tires and Tubes	41
2	Dalton	Perez	Tires and Tubes	40
3	Jennifer	Simmons	Tires and Tubes	38
4	Ashley	Henderson	Tires and Tubes	38
5	Henry	Garcia	Tires and Tubes	38
6	Charles	Jackson	Tires and Tubes	37
7	Fernando	Bames	Tires and Tubes	36
8	Mason	Roberts	Tires and Tubes	36
9	Nancy	Chapman	Tires and Tubes	36
10	Ryan	Thompson	Tires and Tubes	36

For the following steps in this lab, please create a table with this SQL-Code:

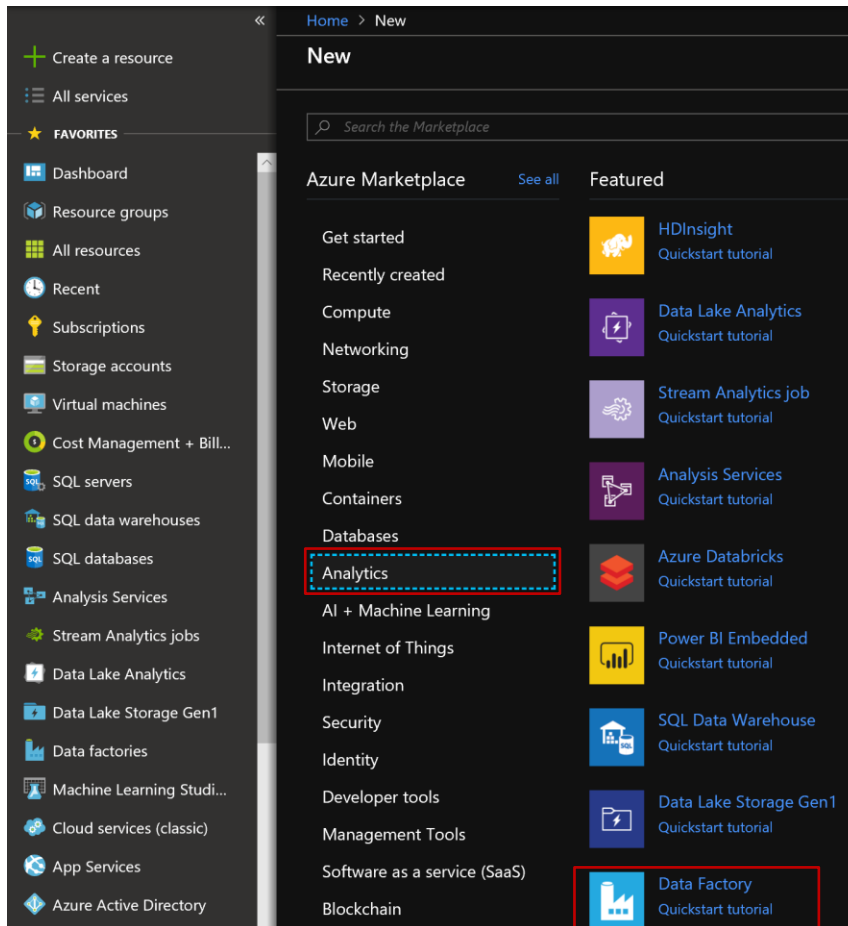
```

CREATE TABLE [dbo].[DimAirports]
(
    [AirportID] [int] NULL,
    [Airport] [nvarchar](100) NULL,
    [City] [nvarchar](100) NULL,
    [Country] [nvarchar](100) NULL,
    [IATA] [nvarchar](100) NULL,
    [ICAO] [nvarchar](100) NULL,
    [Latitude] [nvarchar](100) NULL,
    [Longitude] [nvarchar](100) NULL,
    [Altitude] [nvarchar](100) NULL,
    [Timezone] [nvarchar](100) NULL,
    [DST] [nvarchar](100) NULL,
    [Tz] [nvarchar](100) NULL,
    [AirportType] [nvarchar](100) NULL,
    [DataSource] [nvarchar](100) NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
)
GO

```

## CREATE AN AZURE DATA FACTORY AND LOAD A FILE FROM THE INTERNET INTO OUR SQL DATA WAREHOUSE

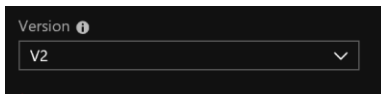
1. Go to your Azure portal and click +, choose Analytics then choose Data Factory



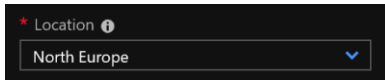
2. Enter a unique name like 'holbigataadf'
3. Choose the resource group you have created in the last sequence for your Data Lake Analytics and Store Services:

The screenshot shows the 'New data factory' form in the Azure portal. The form has the following fields: 'Name' with the value 'holbigataadf' and a green checkmark; 'Subscription' with the value 'Microsoft Azure Internal Consumption (bf2)'; 'Resource Group' with the value 'DataLabsDataFactory' and radio buttons for 'Create new' and 'Use existing' (selected); 'Version' with the value 'V2'; and 'Location' with the value 'East US'. The form is titled 'New data factory' and has a close button in the top right corner.

4. Choose Version: 'V2'



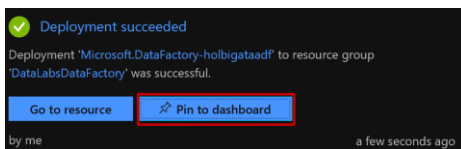
5. Choose the location 'West Europe' or the corresponding location, where you have created your Azure SQL Data Warehouse



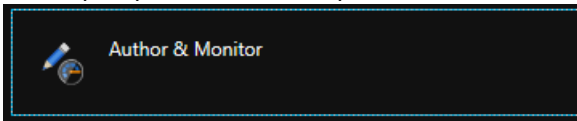
6. Hit 'create' and wait for the service to be displayed. After creation, the Bell Sign will show new messages:



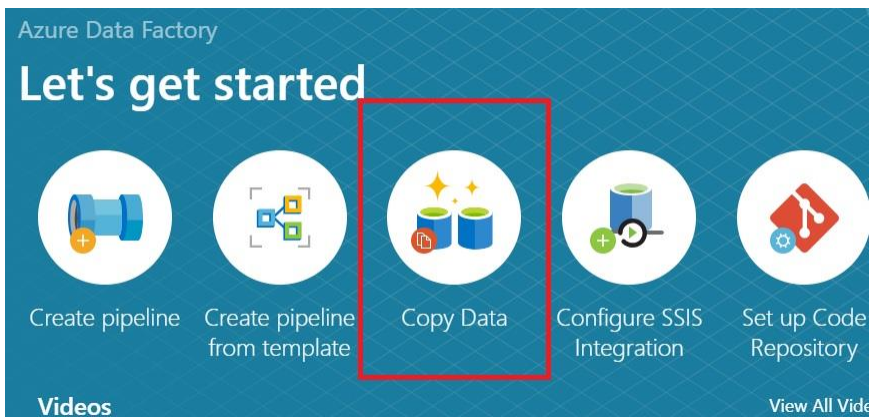
7. Click on it, choose the Deployment-Success message for your Data Factory and click 'Pin to dashboard'



8. 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':



9. On the first screen name your data copy pipeline like 'holbigdatagetdimension' or just leave the default name and click 'Next'



**Copy Data**

- 1 Properties
- 2 Source
  - Connection
  - Dataset
- 3 Destination
  - Connection
  - Dataset
- 4 Settings
- 5 Summary
- 6 Deployment

### Properties

Enter name and description for the copy data task.

Task name \*

Task description

Task cadence or Task schedule

☒ Run once now ☐ Run regularly on schedule

Previous Next

10. On the Source data store screen hit '+ Create new connection':

**Copy Data**

- 1 Properties
  - One time copy
- 2 Source
  - Connection
  - Dataset

### Source data store










Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

All Azure Database File Generic Protocol NoSQL Services and apps

All  + Create new connection

11. Now the 'New Linked Service' picker appears. Please choose "Http" (at the top of the window you even can search for this) and hit 'Continue'

**New Linked Service** X

GE Historian	Google AdWords (Preview)	Google BigQuery
 Greenplum	 HBase	 HDFS
 HTTP	 Hive	 HubSpot (Preview)
 Informix	 Jira (Preview)	 Magento (Preview)

Cancel Continue

12. On the following dialog name the Linked Service and then please enter this URL in the field "Base URL": <https://raw.githubusercontent.com/lzurcher/BigDataHoL/master/airports.dat>  
In the "Authentication type" field select "Anonymous"

← New Linked Service (HTTP) ×

Name \*  
Airportdata

Description

Connect via integration runtime \* ⓘ  
AutoResolveIntegrationRuntime

Base URL \*  
https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat

Server Certificate Validation ⓘ  
Enable

Authentication type \*  
Anonymous

Annotations  
+ New | Delete

□ NAME

Connection successful

Cancel Test connection Finish

Test your connection. If it says "Connection successful" you can click "Finish" and proceed to the next step.

13. On the "Specify HTTP dataset properties" you don't need to change settings. Please just proceed with "Next"

Specify HTTP dataset properties

Relative Url

Request Method  
GET

Additional Headers

☐ Binary Copy ⓘ

Compression Type  
None

Request timeout  
00:01:40

Previous Next

14. In the "File format settings" please choose "Text format" in the "File format" property and hit "Detect Text Format" to let the service decide what column delimiter, linebreaks, etc. are used.

Otherwise you might enter these values yourself. Hit “Next”

### File format settings

File format ?  
Text format ▼ [Detect Text Format](#)

Column delimiter  
Comma (,) ▼  
☐ Use custom delimiter

Row delimiter  
Line feed (\n) ▼  
☐ Use custom delimiter

Skip line count ?  
0

☐ Column names in the first row

▶ Advanced

Preview							
Prop_0	Prop_1	Prop_2	Prop_3	Prop_4	Prop_5	Prop_6	Prop_7
1	"Goroka Airport"	"Goroka"	"Papua New Guinea"	"GKA"	"AYGA"	-6.081689834590001	145.34
2	"Madang Airport"	"Madang"	"Papua New Guinea"	"MAG"	"AYMD"	-5.20707988739	145.71

[Previous](#) [Next](#)

15. Please click “Next” and create a new connection to your Azure SQL Data Warehouse that you have created in the steps above. ➔ “Create new connection”

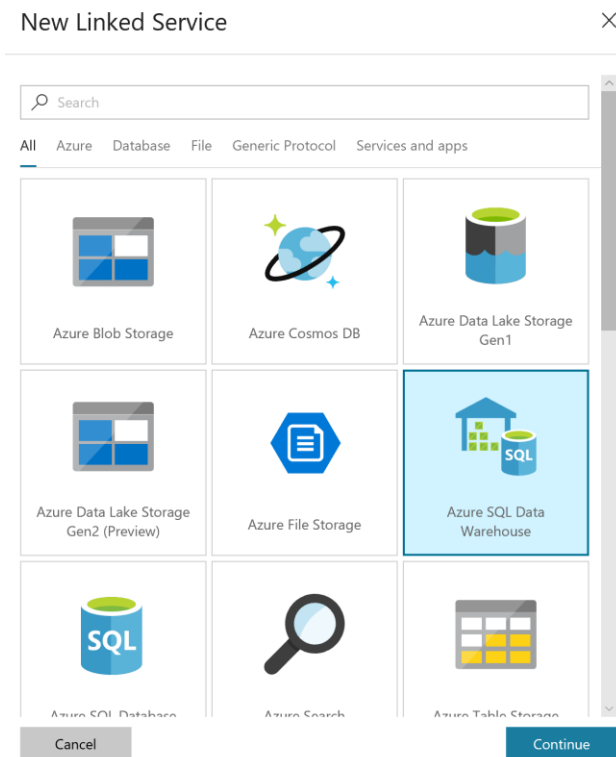
Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

All Azure Database File Generic Protocol NoSQL Services and apps

All ▼  [+ Create new connection](#)

16. Please select "Azure SQL Data Warehouse" and hit "Continue"



17. Name the new LinkedService after your SQL DW or in a manner that you can identify it afterwards. Use "From Azure subscription". Then select your subscription, the SQL DW Server and enter user name and password, that you have created.

Test your connection.

← New Linked Service (Azure SQL)

Name \*  
AzureSqlDW1

Description

Connect via integration runtime \*  
AutoResolveIntegrationRuntime

Connection String

Account selection method  
☒ From Azure subscription ☐ Er

Azure subscription  
Microsoft Azure Internal Consumption (e543a5c5-f7ca-426

Server name \*  
mydatalab

Database name \*  
datalab

Authentication type \*  
SQL Authentication

User name \*  
datalab

Password

Password \*  
\*\*\*\*\*

Additional connection properties

✓ Conf

Cancel Tr

18. On the "Destination Data Store" your new Linked Service shows up and is selected. Hit "Next". You are taken to the "Table mapping" dialog. Select the newly created table that you have created in the former step above.

Table mapping

For each table you have selected to copy in the source data store, select a corresponding table in the destination data store or sp destination.

Source	Destination
HTTP file	→ [dbo].[DimAirports] ↺

☐ Skip column mapping for all tables

Previous Next

19. The “Column mapping” dialog gives you the chance to map the source columns from the file to the target columns in the database table. You can leave this as is. The columns should map in the right order in this case. Hit “Next”.

Column mapping

Choose how source and destination columns are mapped

Table mappings (1)

Source  
HTTP file  
Destination  
[dbo].[DimAirports]

Column mappings

HTTP file [dbo].[DimAirports]

Prop_0 (int64)	→	AirportID (int32)
Prop_1 (String)	→	Airport (String)
Prop_2 (String)	→	City (String)
Prop_3 (String)	→	Country (String)
Prop_4 (String)	→	IATA (String)
Prop_5 (String)	→	ICAO (String)
Prop_6 (String)	→	Latitude (String)

Azure SQL Data Warehouse sink properties

Pre-copy script

Write batch size

Previous Next

20. The “Settings” dialogue will appear and gives you the chance for final adjustments. We want to allow Azure Data Factory and Azure SQL DW to use Polybase to increase the speed when loading into the DB. So therefore please create a new Storage account, that will be used for staging during loads into the target database. Data Factory will do this job for you. So please click “ + New” or if you already have a storage account, select one in the drop down box.

Settings

More options for data movement

Fault tolerance settings

Fault tolerance Abort activity on first incompatible row ⓘ

Performance settings

☒ Enable Staging ⓘ

Staging Settings

Staging Account Linked Service Select... ⓘ + New

Storage Path ⓘ Browse

☐ Enable Compression ⓘ

Advanced settings

☒ Allow polybase ⓘ

Reject type Value

Reject value 0

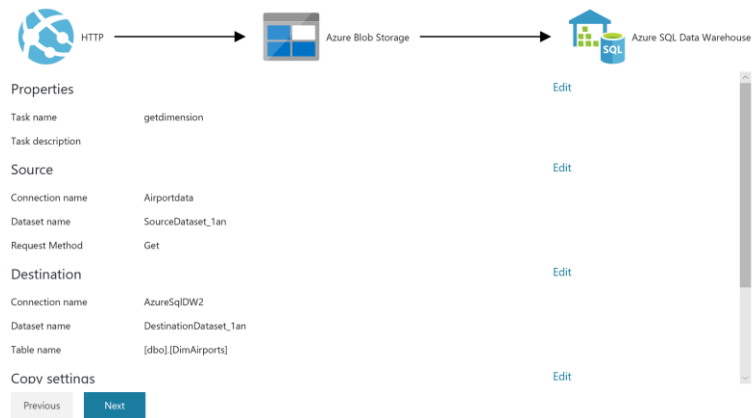
Previous Next

21. In the next and last step, a “Summary” about the newly created pipeline is shown. If everything is setup to your needs and requirements, you might hit “Next” and the pipeline will be deployed

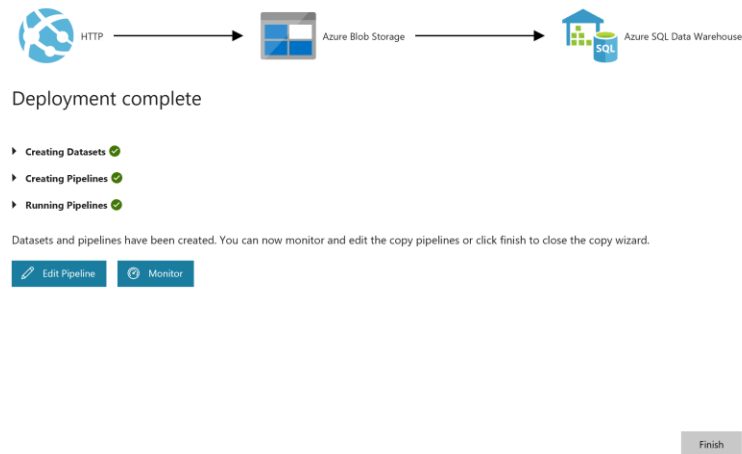
and executed.

### Summary

You are running pipeline to copy data from HTTP to Azure SQL Data Warehouse.



22. You might jump into the monitoring of your Data Factory to collect information about the pipeline run and check, if there were errors in the execution.



23. You could now get back to the SQL Server Management Studio and check the content of the table. (select \* from dbo.DimAirports)

## LAB 2: CREATING AND LOADING A DATA LAKE GEN 2

### OVERVIEW

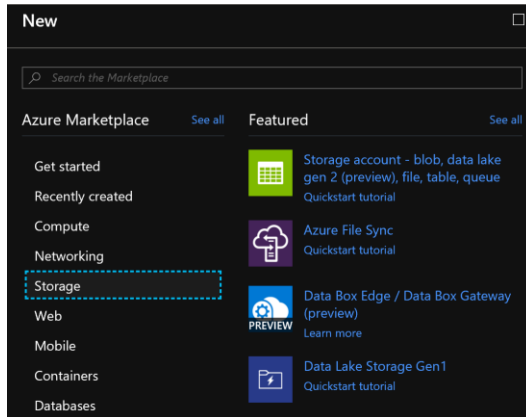
In this Lab you will create a Data Factory Pipeline, that will load data from airdelay statistics into your then newly created Azure Data Lake Gen2.

This data will then be cleansed and pre-aggregated for analysis and Machine Learning using a Databricks Spark-as-a-Service Cluster.

### PRE-LOAD YOUR DATA TO DATA LAKE

First we will create a Storage Account, that will hold your file data.

1. Please go to the portal, click '+ Create a resource' and on the Storage – Tab select 'Storage Account'





- On the next Blade please name Storage Account, choose the resource group that you have created above and choose 'West Europe' as your selected region:

Home > New > Create storage account

### Create storage account

Basics Advanced Tags Review + create

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Az Tables. The cost of your storage account depends on the usage and the options you choose.

PROJECT DETAILS

Select the subscription to manage deployed resources and costs. Use resource groups like f your resources.

\* Subscription: Microsoft Azure Internal Consumption (e543a5c5-...)

\* Resource group: datalabrg [Create new](#)

INSTANCE DETAILS

The default deployment model is Resource Manager, which supports the latest Azure featur the classic deployment model instead. [Choose classic deployment model](#)

\* Storage account name: mydatalabsagen2

\* Location: West Europe

Performance: ☒ Standard ☐ Premium

Account kind: StorageV2 (general purpose v2)

Replication: Read-access geo-redundant storage (RA-GRS)

Access tier (default): ☐ Cool ☒ Hot

- In the Advanced tab make sure you check Hierarchical namespace "Enabled" to enable a Data Lake Storage Gen2

Home > New > Create storage account

### Create storage account

Basics Advanced Tags Review + create

SECURITY

Secure transfer required: ☐ Disabled ☒ Enabled

VIRTUAL NETWORKS

Allow access from: ☒ All networks ☐ Selected network

DATA PROTECTION

Block soft delete: ☐ Disabled ☒ Enabled

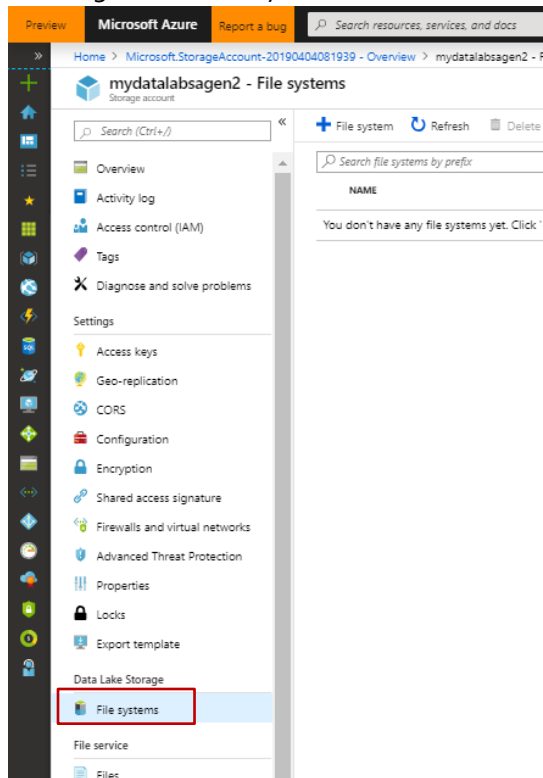
DATA LAKE STORAGE GEN2

Hierarchical namespace: ☐ Disabled ☒ Enabled

[Review + create](#) [Previous](#) [Next: Tags >](#)

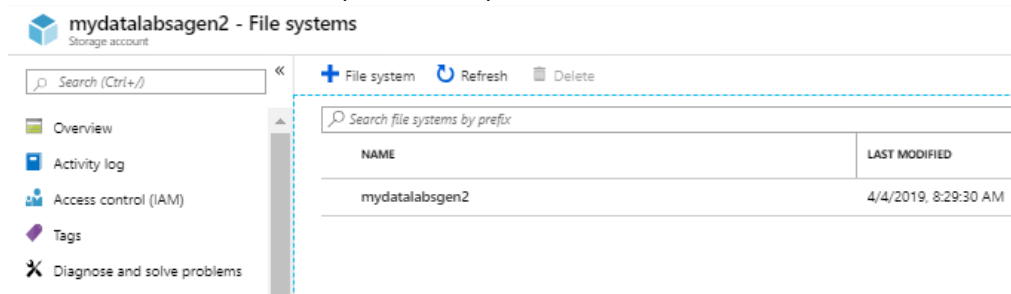
- When you hit create it will take some minutes and the service will appear available in your resource group.

5. Please go to the newly created Data Lake Storage Gen 2 and hit 'File systems'

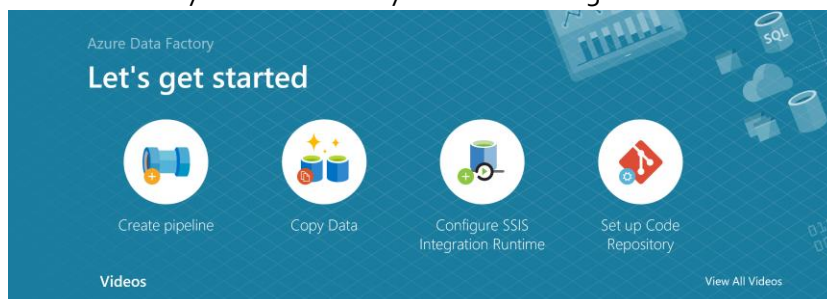


6. Add file system

7. Create a new folder to hold your airdelay-files:



8. Now switch to your Data Factory environment again and start a new 'Copy Data' Wizard:



9. On the first screen name your data copy pipeline like 'holbigdataairdelay' or just leave the default name and click 'Next'

The screenshot shows the 'Copy Data' task configuration interface. On the left, a sidebar lists the steps: 1 Properties, 2 Source, 3 Destination, 4 Settings, 5 Summary, and 6 Deployment. The 'Properties' step is selected. The main area is titled 'Properties' and contains the following fields: 'Task name' (with the value 'GetImages'), 'Task description' (empty), and 'Task cadence or Task schedule' (with 'Run once now' selected). At the bottom, there are 'Previous' and 'Next' buttons.

10. On the Source data store screen hit '+ Create new connection':

The screenshot shows the 'Source data store' configuration screen. The left sidebar shows 'Source' selected, with 'Connection' and 'Dataset' options. The main area is titled 'Source data store' and contains a tabbed interface with 'All', 'Azure', 'Database', 'File', 'Generic Protocol', 'NoSQL', and 'Services and apps'. Below the tabs, there is a dropdown menu set to 'All', a 'Filter by name' input field, and a red-bordered button labeled '+ Create new connection'.

11. Now the 'New Linked Service' picker appears. Please choose Azure Blob Storage and hit 'Continue'

The screenshot shows the 'New Linked Service' picker dialog. It has a search bar at the top and a grid of service icons. The services listed are: Amazon Marketplace Web Service (Preview), Amazon Redshift, Amazon S3, Apache Impala (Preview), Azure Blob Storage (highlighted with a red border), Azure Cosmos DB, Azure Data Lake Storage, Azure Data Lake Storage, and Azure Database for MySQL. At the bottom, there are 'Cancel' and 'Continue' buttons.

12. On the following screen select 'Use SAS URI' as 'Authentication Method' and paste the following into the SAS URI-Field:

← New Linked Service (Azure Blob Storage) ×

Name \*  
AirdelaySourcedata

Description

Connect via integration runtime \*  
AutoResolveIntegrationRuntime

Authentication method  
SAS URI

SAS URI  
Azure Key Vault

SAS URL \*  
sample: https://myaccount.blob.core.windows.net/sascontainer/sasblob.txt

SAS Token  
Azure Key Vault

SAS Token  
sample: ?sv=<storage services version>&st=<start time>&se=<expire time>&sr=<resource>&sp=<per>

Annotations  
+ New

▶ Advanced ⓘ

### SAS URL

<https://mydatalabstorage.blob.core.windows.net/>

### SAS Token

?sv=2018-03-28&ss=bfqt&srt=sco&sp=rwdlacup&se=2019-04-04T14:35:44Z&st=2019-04-04T06:35:44Z&spr=https&sig=FXycESAAPNgPrP6fdpDXPm78l8Hx1Hp9o%2BOFOW4e2hY%3D

The 'Choose input file or folder' selection is displayed. First click the checkbox 'Binary Copy', then please select 'Browse'

13. On the following screen choose the folder 'airdelays' (and really click on 'Choose' to make the section active. Then click 'Next'

Choose the input file or folder

Select a source folder or file to be copied to the destination data store.

File or folder \*  
adfstagedpolybasetempdata/ ⓘ Browse

↑ ↗ >

adfstagedpolybasetempdata

14. On the 'Destination data store' dialogue we again create another connection:

### Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

All Azure Database File Generic Protocol NoSQL Services and apps


All Filter by name + Create new connection

15. Please select 'Azure Data Lake Storage Gen2' and hit 'Continue'


## New Linked Service

Search


All Azure File



Azure Blob Storage



Azure Data Lake Storage Gen1



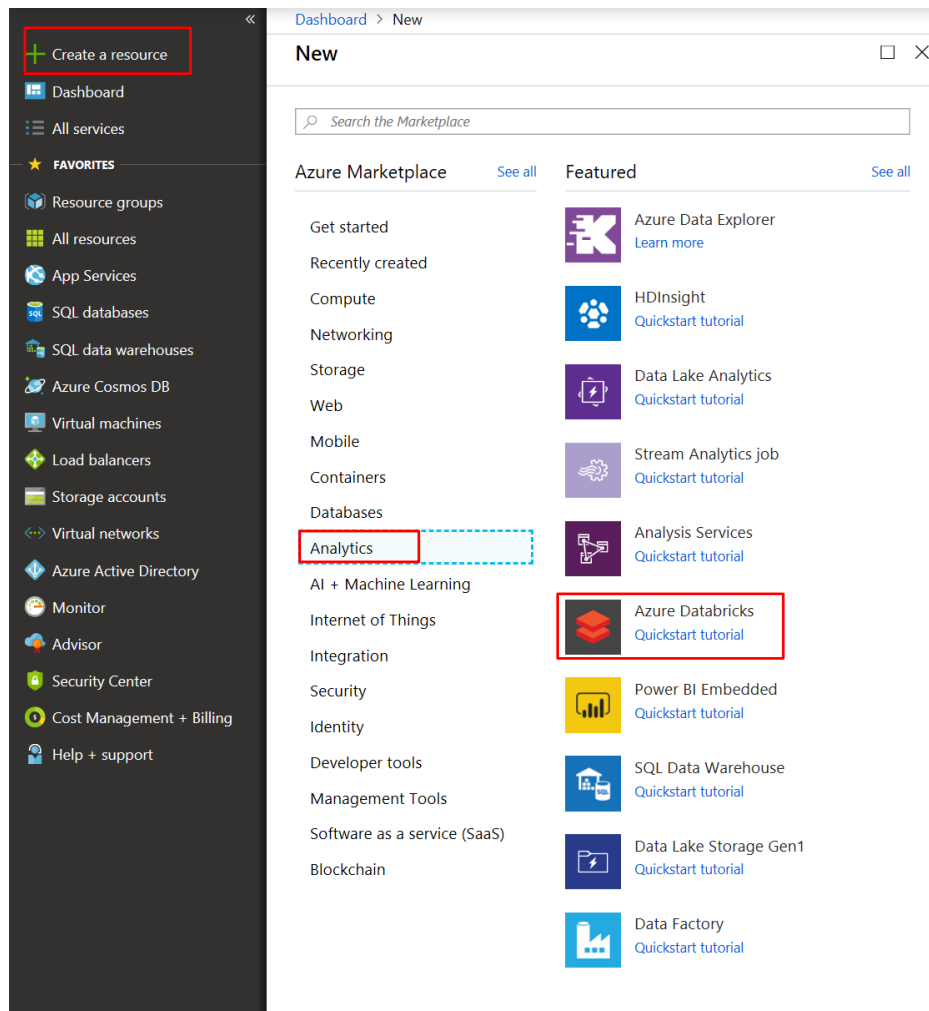
Azure Data Lake Storage Gen2

16. On the following dialogue you may name the connection to a name of your choice and then select the following entries:
- Select your Azure Subscription from the dropdown box
  - select the Data Lake Storage Gen2 account name that you have created above
- Then hit 'Finish'

## LAB 3 CONNECT WITH STORAGE EXPLORER TO SEE THE STORAGE ACCOUNTS.

Let's create a Databricks workspace to manipulate and analyze the data using Spark.

Go in the Azure Portal and create a new Azure Databricks resource.



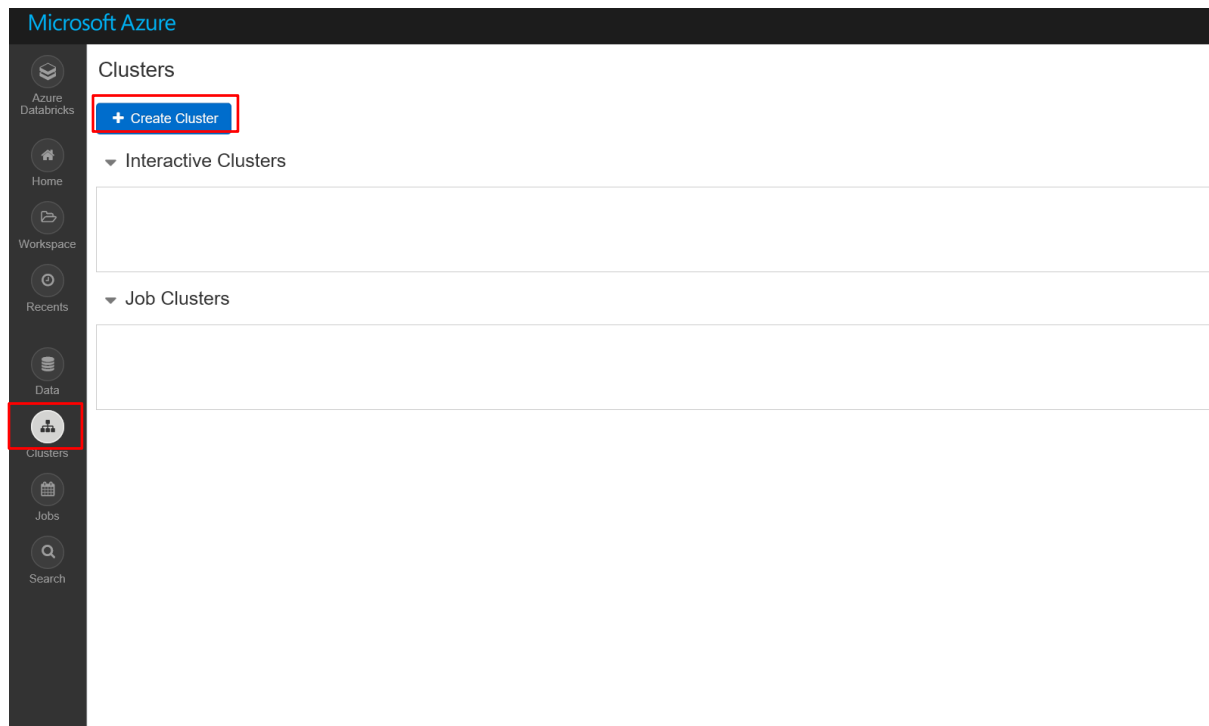
Configure it with:

- A name
- Your subscription
- The resource group you previously created
- Location: North Europe
- Choose the Standard pricing tier

When the resource is created, click on *Launch Workspace*.



Go in the *Clusters* tab and click on *Create Cluster*.



## Create Cluster

### New Cluster

Cancel

Create Cluster

2-8 Workers: 28.0-112.0 GB Memory, 8-32 Cores, 1.5-6 DBU  
1 Driver: 14.0 GB Memory, 4 Cores, 0.75 DBU Cost \$0.40 per DBU ⓘ

#### Cluster Name

ClusterLab

#### Cluster Mode

☐ High Concurrency

Optimized to run concurrent SQL, Python, and R workloads.  
Does not support Scala. Previously known as Serverless.

☒ Standard

Recommended for single-user clusters. Can run SQL, Python, R,  
and Scala workloads.

#### Databricks Runtime Version ⓘ

4.3 (includes Apache Spark 2.3.1, Scala 2.11)

#### Python Version ⓘ

2

#### Driver Type

Same as worker

14.0 GB Memory, 4 Cores, 0.75 DBU

#### Worker Type

Standard\_DS3\_v2

14.0 GB Memory, 4 Cores, 0.75 DBU

Min Workers

2

Max Workers

8

☒ Enable autoscaling ⓘ

#### Auto Termination ⓘ

☒ Terminate after 120 minutes of inactivity

Spark

Tags

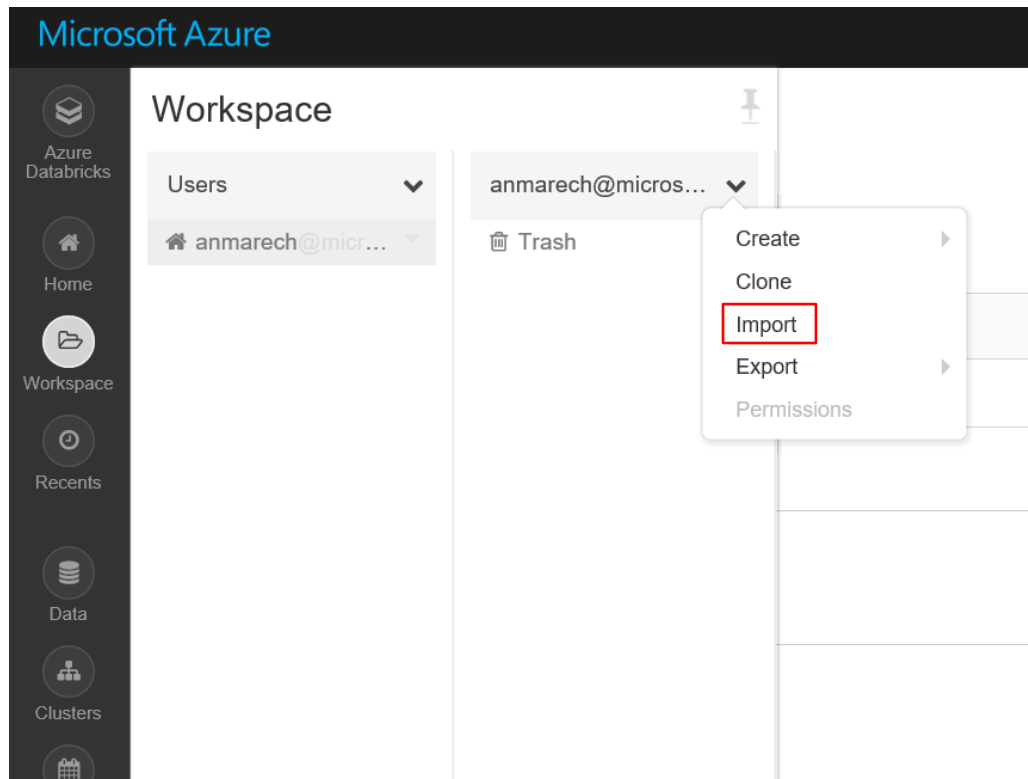
Logging

Init Scripts



Give a name to your cluster. You can leave the other properties as they are.

Once your cluster preparation is running, explore the Databricks environment and import the notebooks in your workspace.



For the next steps, let's run the Databricks notebook.

## LAB 4

Can you create a Cosmos DB using SQL API and insert the json file from

<https://github.com/lzurcher/BigDataHol> ? Once inserted, copy the document and insert a new one.

Add the property "my\_favorite\_song" to the file and give it the value of your favorite song. 😊

## TERMS OF USE

© 2017 Microsoft Corporation. All rights reserved.

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 FUNCTIONALITY 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 DataBricks and Azure Data Lake Storage. Some of the features might change in future releases of the product.