Microsoft Tech Series Lab: Data Lifecycle part 2

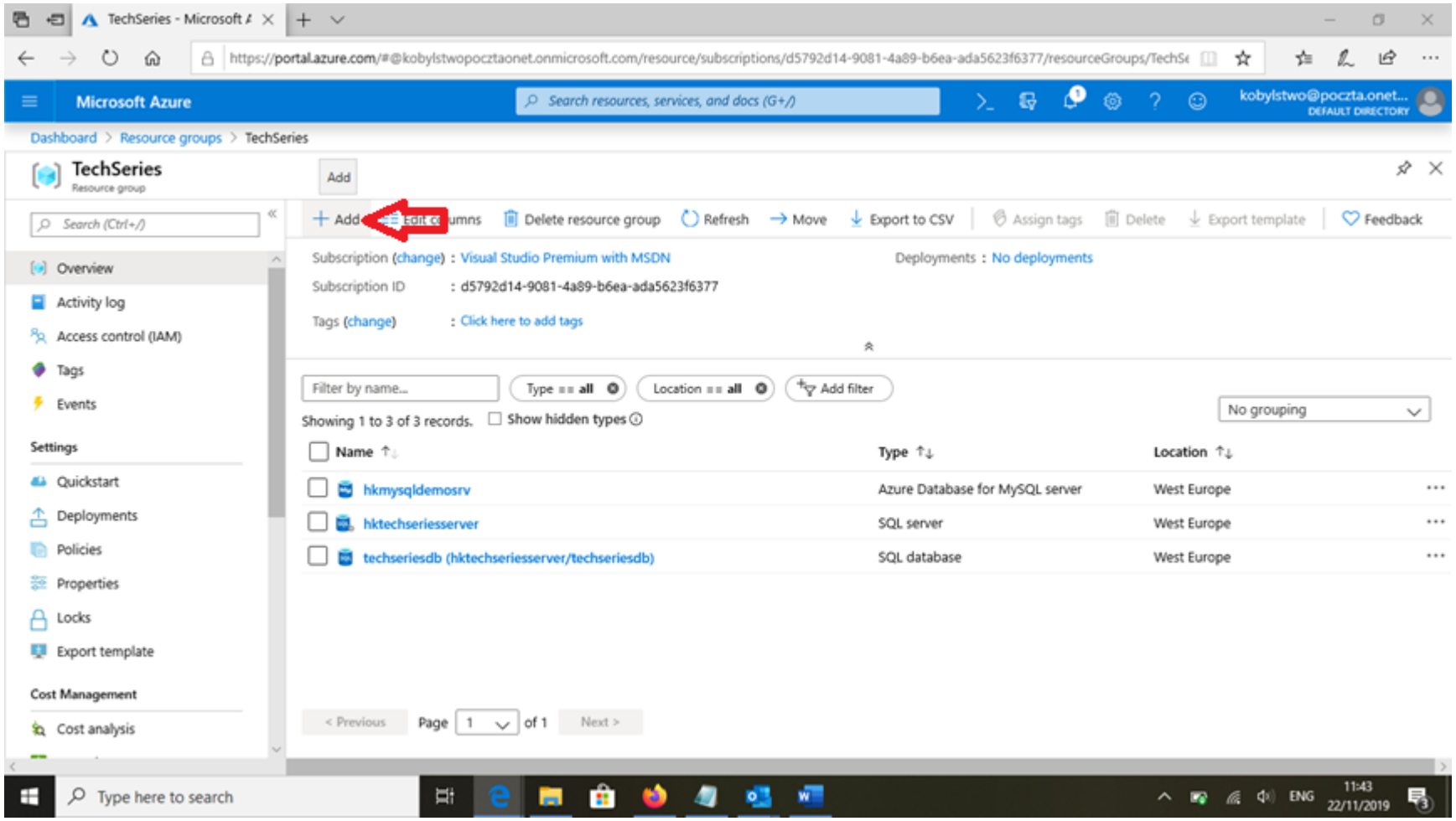
by Hubert K. Kobierzewski from Data Community Poland

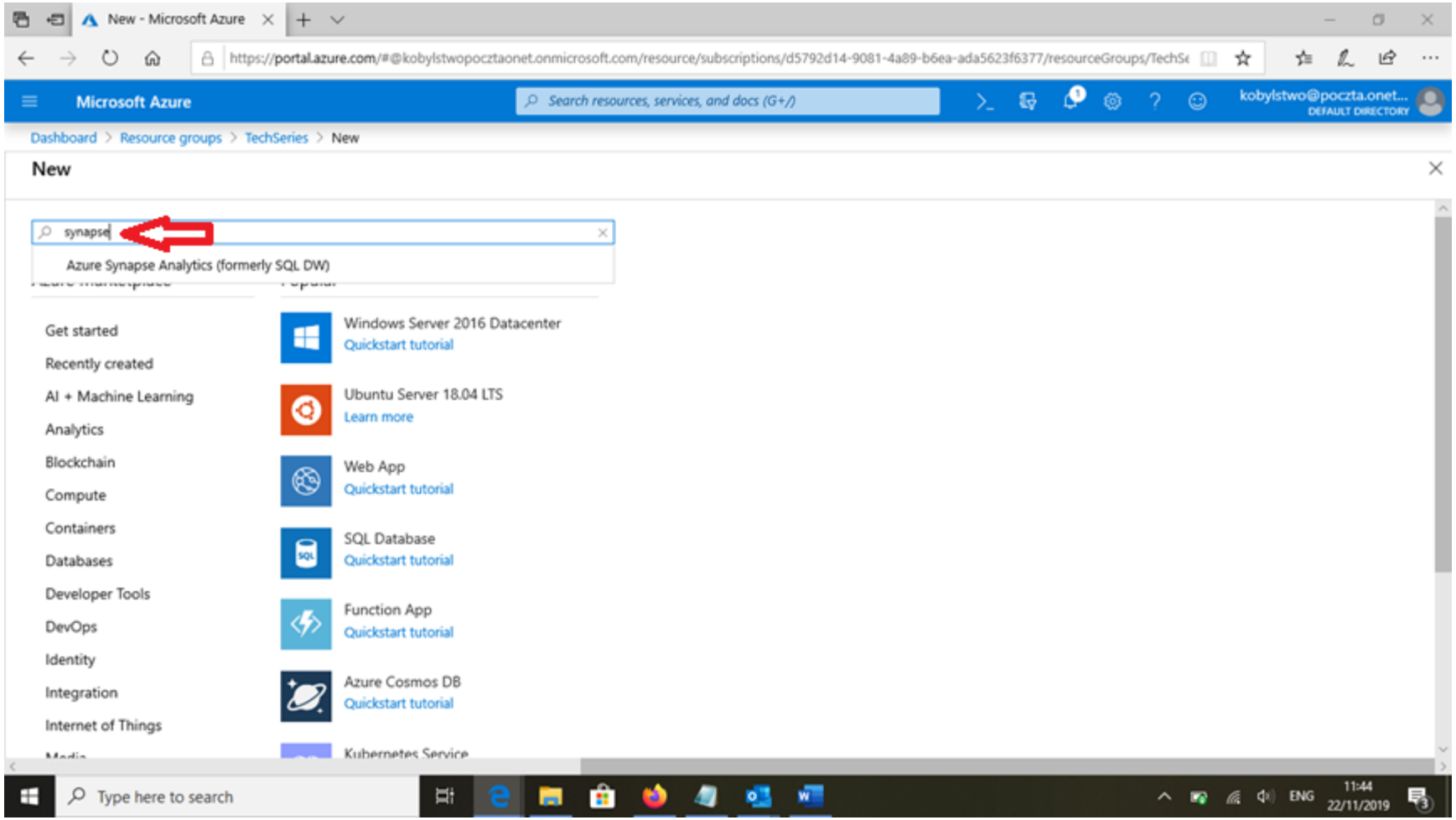
The following scenario was covered during Microsoft Tech Series event on Nov 22nd, 2019

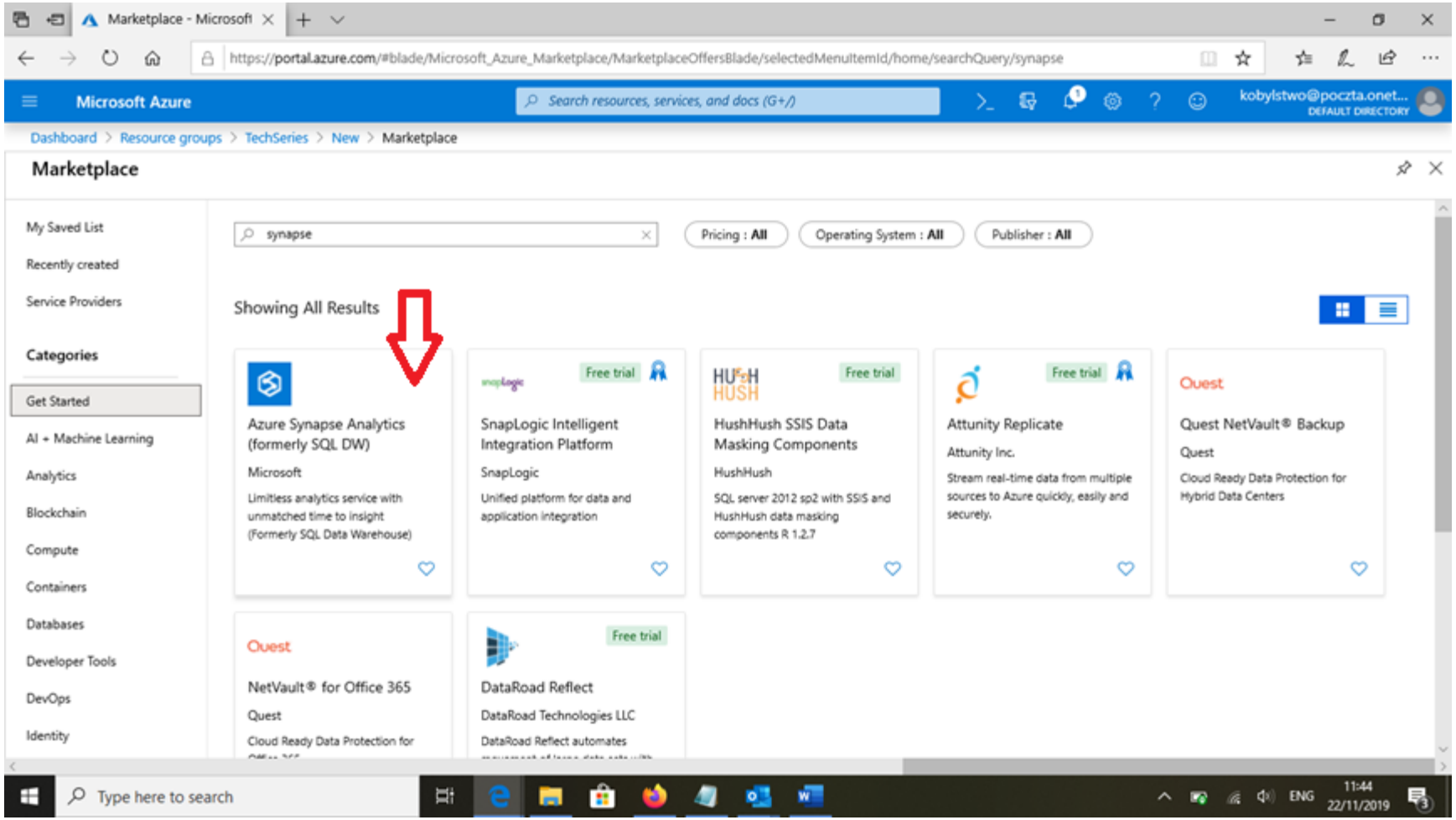
(Before you start, follow all steps provided by Damian Widera https://tiny.pl/tnpbn)

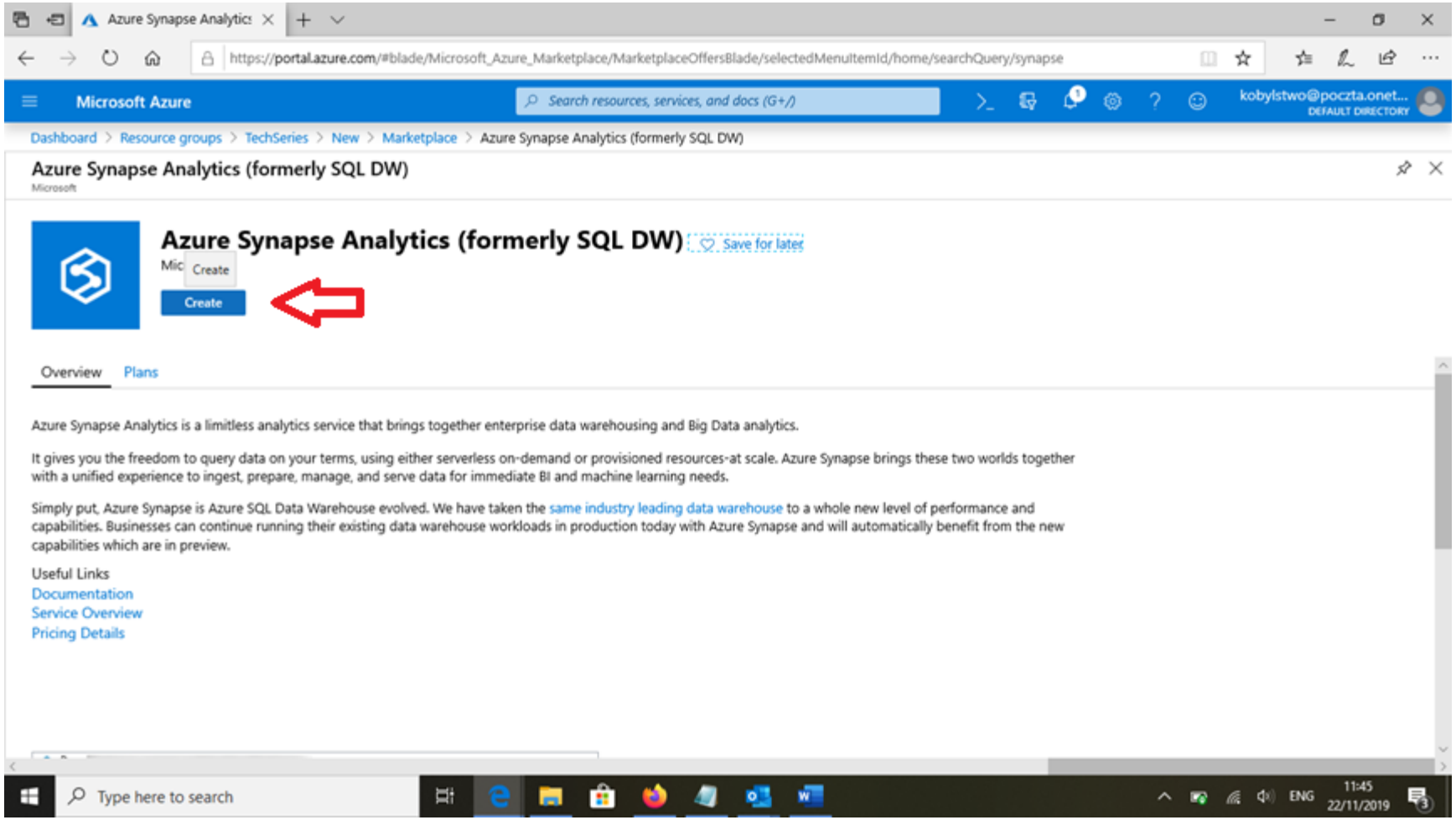
Create Azure Synapse Analytics instance (aka SQL DW). We will need it later on.

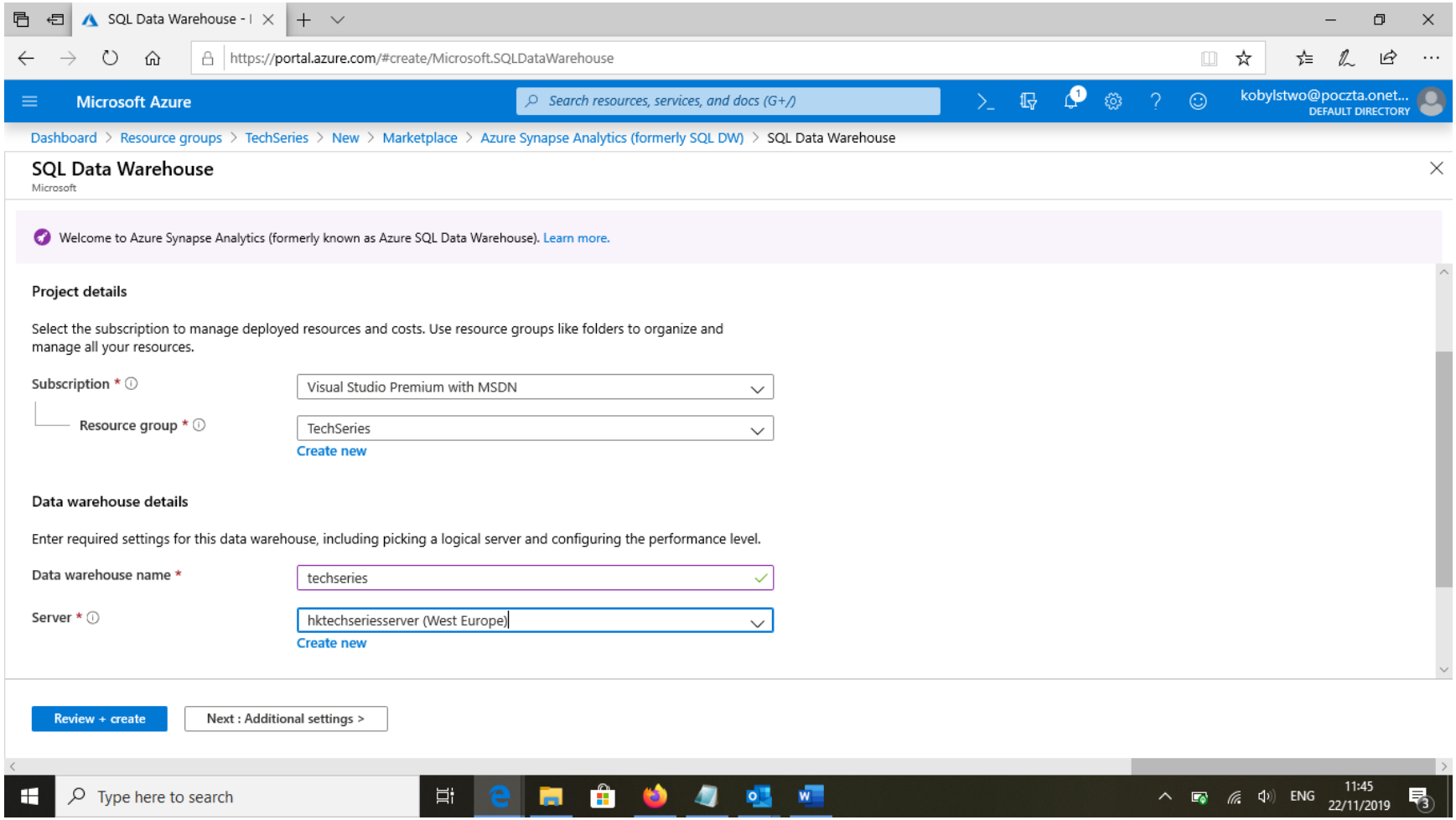
Use existing virtual SQL Server – this will make our exercise easier.

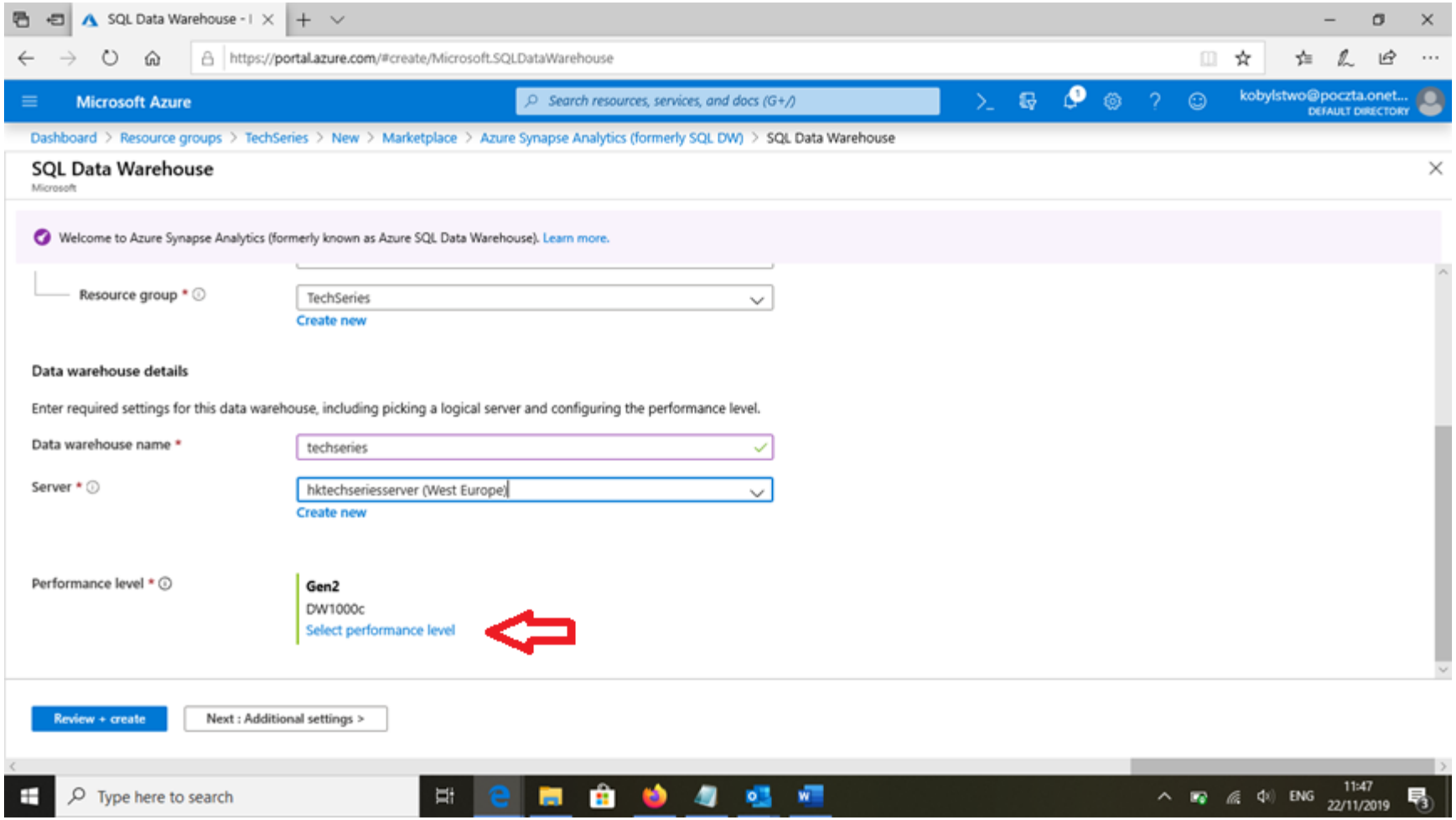


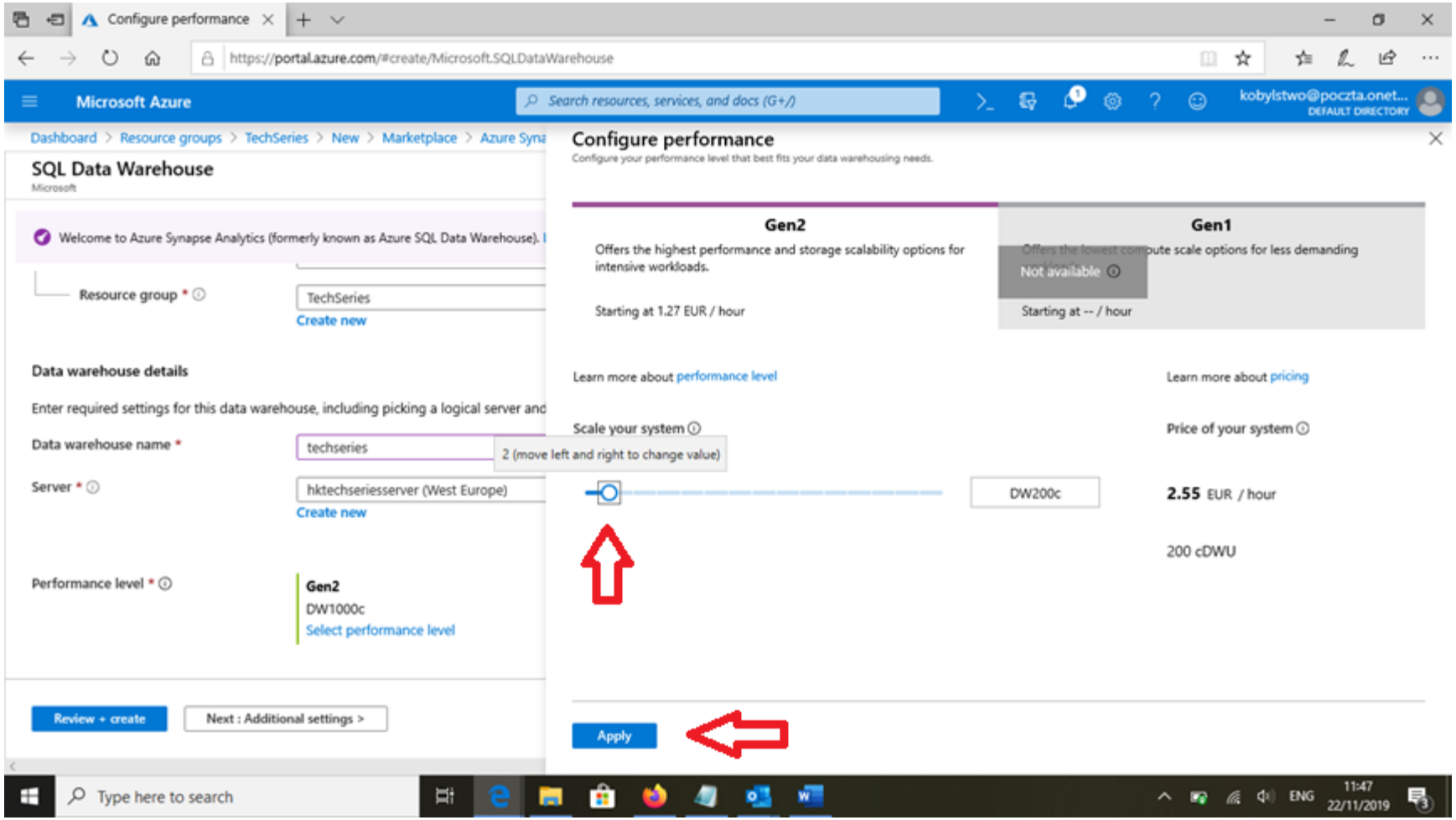

New - Microsoft Azure
https://portal.azure.com/#@kobylstwopocztaonet.onmicrosoft.com/resource/subscriptions/d5792d14-9081-4a89-b6ea-ada5623f6377/resourceGroups/TechSe
Microsoft Azure
Search resources, services, and docs (G+/)
kobylstwo@poczta.onet...
DEFAULT DIRECTORY
Dashboard > Resource groups > TechSeries > New
New
synapse
Azure Synapse Analytics (formerly SQL DW)
Get started
Recently created
AI + Machine Learning
Analytics
Blockchain
Compute
Containers
Databases
Developer Tools
DevOps
Identity
Integration
Internet of Things
Windows Server 2016 Datacenter
Quickstart tutorial
Ubuntu Server 18.04 LTS
Learn more
Web App
Quickstart tutorial
SQL Database
Quickstart tutorial
Function App
Quickstart tutorial
Azure Cosmos DB
Quickstart tutorial
Kubernetes Service
Type here to search
ENG
11:44
22/11/2019

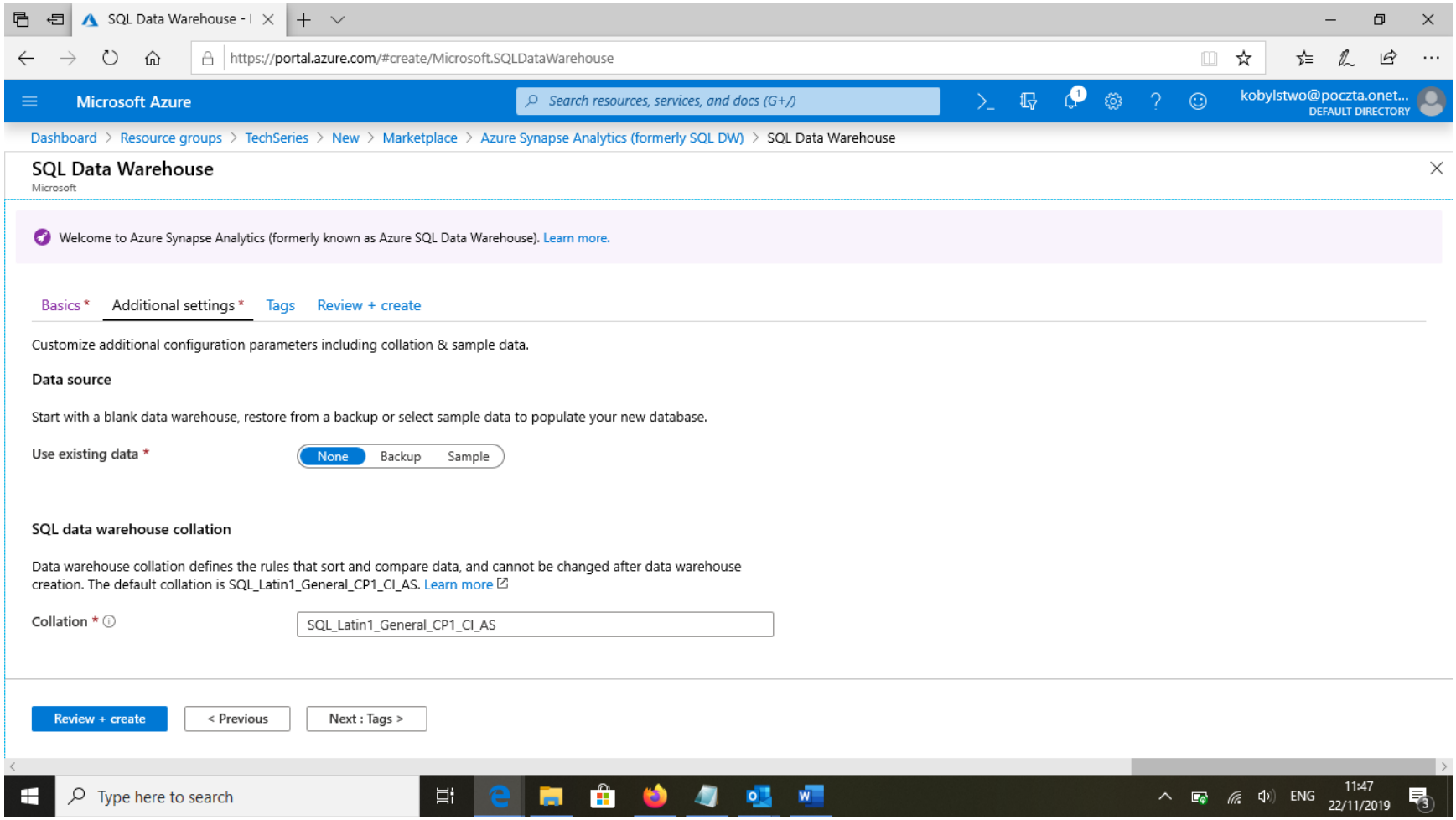

Marketplace - Microsoft
https://portal.azure.com/#blade/Microsoft_Azure_Marketplace/MarketplaceOffersBlade/selectedMenuItemId/home/searchQuery/synapse
Microsoft Azure
Search resources, services, and docs (G+/)
kobylstwo@poczta.onet...
DEFAULT DIRECTORY
Dashboard > Resource groups > TechSeries > New > Marketplace
Marketplace
My Saved List
Recently created
Service Providers
Categories
Get Started
AI + Machine Learning
Analytics
Blockchain
Compute
Containers
Databases
Developer Tools
DevOps
Identity
synapse
Pricing : All
Operating System : All
Publisher : All
Showing All Results
Azure Synapse Analytics (formerly SQL DW)
Microsoft
Limitless analytics service with unmatched time to insight (Formerly SQL Data Warehouse)
Free trial
SnapLogic Intelligent Integration Platform
SnapLogic
Unified platform for data and application integration
Free trial
HushHush SSIS Data Masking Components
HushHush
SQL server 2012 sp2 with SSIS and HushHush data masking components R 1.2.7
Free trial
Attunity Replicate
Attunity Inc.
Stream real-time data from multiple sources to Azure quickly, easily and securely.
Quest
Quest NetVault® Backup
Quest
Cloud Ready Data Protection for Hybrid Data Centers
Quest
NetVault® for Office 365
Quest
Cloud Ready Data Protection for
Free trial
DataRoad Reflect
DataRoad Technologies LLC
DataRoad Reflect automates
Type here to search
ENG
11:44
22/11/2019

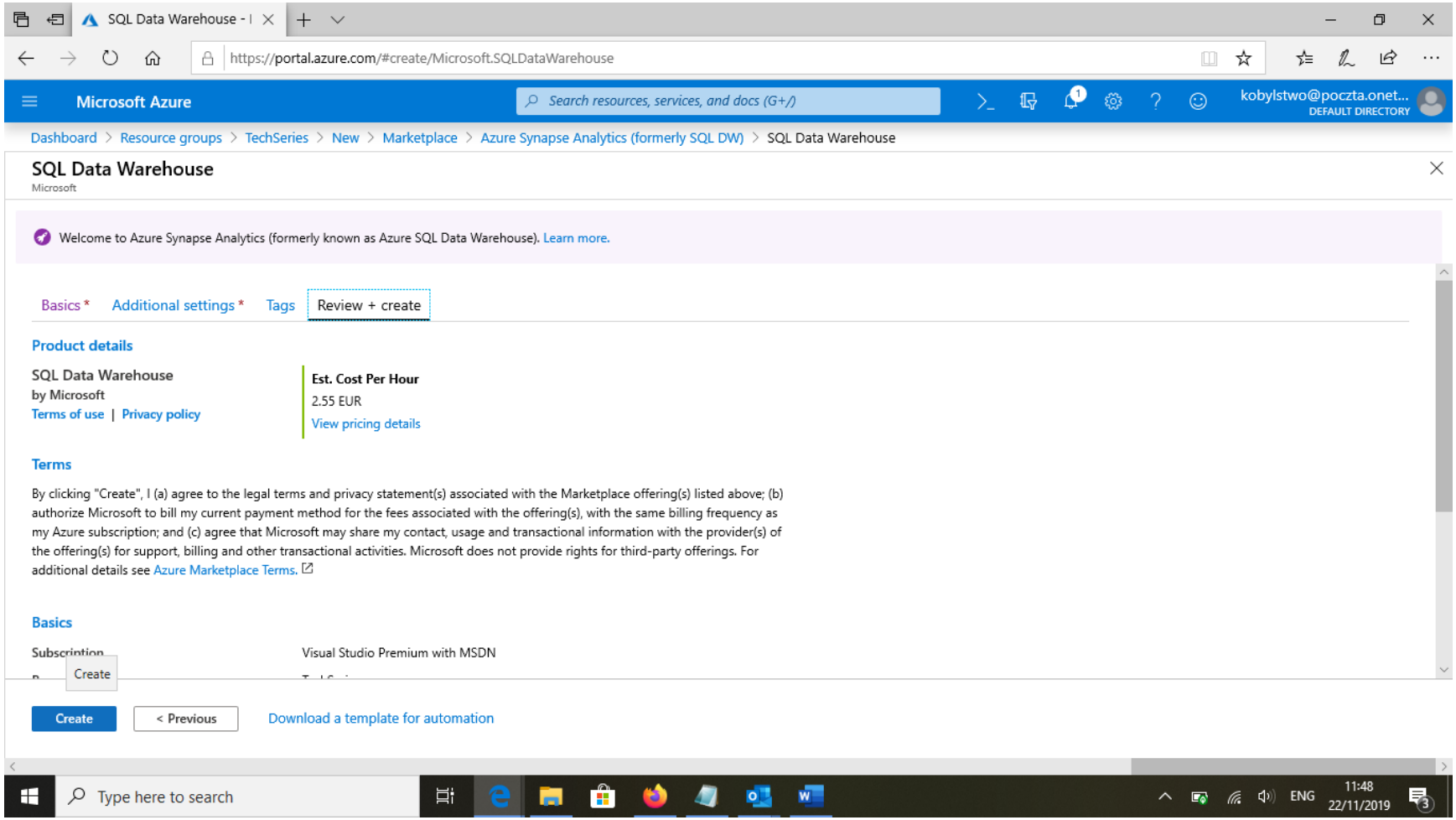

Azure Synapse Analytics
https://portal.azure.com/#blade/Microsoft_Azure_Marketplace/MarketplaceOffersBlade/selectedMenuItemId/home/searchQuery/synapse
Microsoft Azure
Search resources, services, and docs (G+/)
kobylstwo@poczta.onet...
DEFAULT DIRECTORY
Dashboard > Resource groups > TechSeries > New > Marketplace > Azure Synapse Analytics (formerly SQL DW)
Azure Synapse Analytics (formerly SQL DW)
Microsoft
Azure Synapse Analytics (formerly SQL DW)
Save for later
Create
Create
Overview
Plans
Azure Synapse Analytics is a limitless analytics service that brings together enterprise data warehousing and Big Data analytics.
It gives you the freedom to query data on your terms, using either serverless on-demand or provisioned resources-at scale. Azure Synapse brings these two worlds together with a unified experience to ingest, prepare, manage, and serve data for immediate BI and machine learning needs.
Simply put, Azure Synapse is Azure SQL Data Warehouse evolved. We have taken the same industry leading data warehouse to a whole new level of performance and capabilities. Businesses can continue running their existing data warehouse workloads in production today with Azure Synapse and will automatically benefit from the new capabilities which are in preview.
Useful Links
Documentation
Service Overview
Pricing Details
Type here to search
ENG
11:45
22/11/2019


SQL Data Warehouse - I
https://portal.azure.com/#create/Microsoft.SQLDataWarehouse
Microsoft Azure
Search resources, services, and docs (G+/)
kobylstwo@poczta.onet...
DEFAULT DIRECTORY
Dashboard > Resource groups > TechSeries > New > Marketplace > Azure Synapse Analytics (formerly SQL DW) > SQL Data Warehouse
SQL Data Warehouse
Microsoft
Welcome to Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse). Learn more.
Project details
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.
Subscription * Visual Studio Premium with MSDN
Resource group * TechSeries
Create new
Data warehouse details
Enter required settings for this data warehouse, including picking a logical server and configuring the performance level.
Data warehouse name * techseries
Server * hktechseriesserver (West Europe)
Create new
Review + create
Next : Additional settings >
Type here to search
ENG
11:45
22/11/2019


SQL Data Warehouse - I
https://portal.azure.com/#create/Microsoft.SQLDataWarehouse
Microsoft Azure
Search resources, services, and docs (G+/)
kobylstwo@poczta.onet...
DEFAULT DIRECTORY
Dashboard > Resource groups > TechSeries > New > Marketplace > Azure Synapse Analytics (formerly SQL DW) > SQL Data Warehouse
SQL Data Warehouse
Microsoft
Welcome to Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse). Learn more.
Resource group *
TechSeries
Create new
Data warehouse details
Enter required settings for this data warehouse, including picking a logical server and configuring the performance level.
Data warehouse name *
techseries
Server *
hktechseriesserver (West Europe)
Create new
Performance level *
Gen2
DW1000c
Select performance level
Review + create
Next : Additional settings >
Type here to search
ENG
11:47
22/11/2019


Configure performance
https://portal.azure.com/#create/Microsoft.SQLDataWarehouse
Microsoft Azure
Search resources, services, and docs (G+/)
kobylstwo@poczta.onet...
DEFAULT DIRECTORY
Dashboard > Resource groups > TechSeries > New > Marketplace > Azure Syna
SQL Data Warehouse
Microsoft
Welcome to Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse).
Resource group *
TechSeries
Create new
Data warehouse details
Enter required settings for this data warehouse, including picking a logical server and
Data warehouse name *
techseries
Server *
hktechseriesserver (West Europe)
Create new
Performance level *
Gen2
DW1000c
Select performance level
Review + create
Next : Additional settings >
Configure performance
Configure your performance level that best fits your data warehousing needs.
Gen2
Offers the highest performance and storage scalability options for intensive workloads.
Starting at 1.27 EUR / hour
Gen1
Offers the lowest compute scale options for less demanding
Not available
Starting at -- / hour
Learn more about performance level
Learn more about pricing
Scale your system
2 (move left and right to change value)
DW200c
Price of your system
2.55 EUR / hour
200 cDWU
Apply
Type here to search
ENG
11:47
22/11/2019

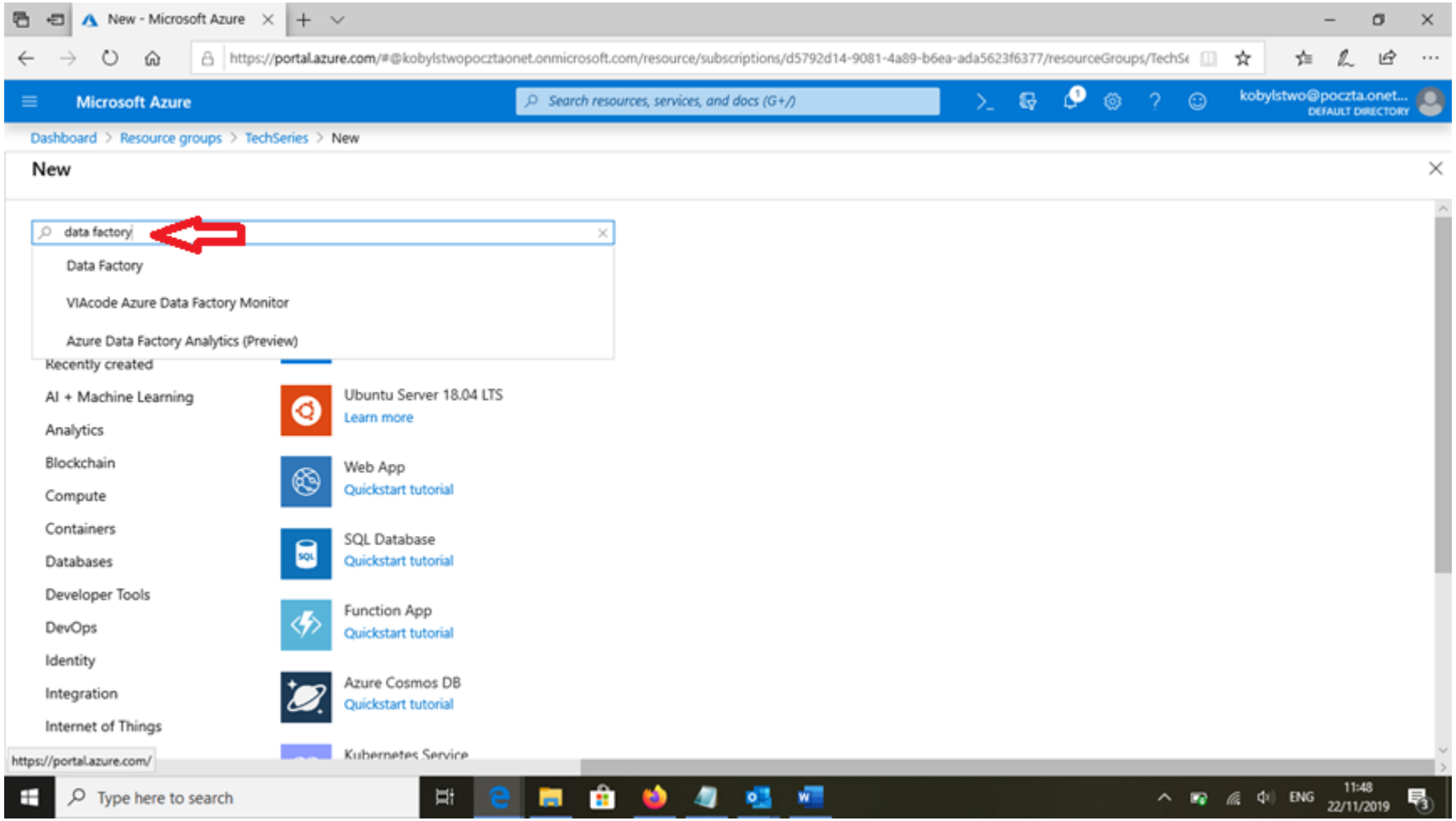

SQL Data Warehouse - I
https://portal.azure.com/#create/Microsoft.SQLDataWarehouse
Microsoft Azure
Search resources, services, and docs (G+/)
kobylstwo@poczta.onet...
DEFAULT DIRECTORY
Dashboard > Resource groups > TechSeries > New > Marketplace > Azure Synapse Analytics (formerly SQL DW) > SQL Data Warehouse
SQL Data Warehouse
Microsoft
Welcome to Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse). Learn more.
Basics * Additional settings * Tags Review + create
Customize additional configuration parameters including collation & sample data.
Data source
Start with a blank data warehouse, restore from a backup or select sample data to populate your new database.
Use existing data * None Backup Sample
SQL data warehouse collation
Data warehouse collation defines the rules that sort and compare data, and cannot be changed after data warehouse creation. The default collation is SQL_Latin1_General_CP1_CI_AS. Learn more
Collation *
SQL_Latin1_General_CP1_CI_AS
Review + create
< Previous
Next : Tags >
Type here to search
ENG
11:47
22/11/2019



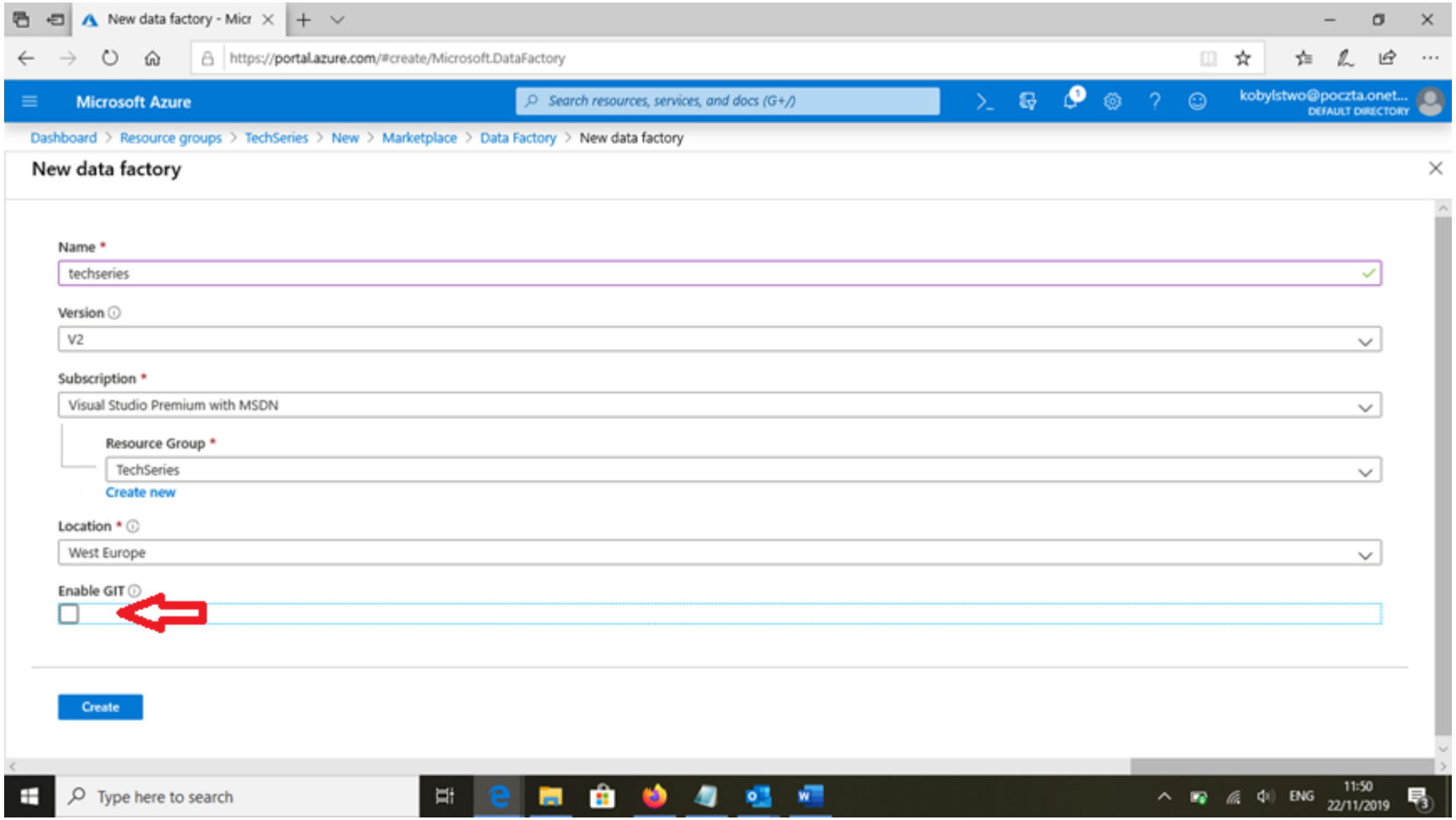
Alternatively you can use the following CLI command:

```
az sql dw create --name techseriesdwh --resource-group TechSeries --server techseriesserver --service-objective DW200c --zone-redundant false
```

Do not wait for creating Azure Synapse Analytics (a.k.a. SQL Data Warehouse) instance, initialize your Data Factory environment. Go to the TechSeries resource group and add a new service.

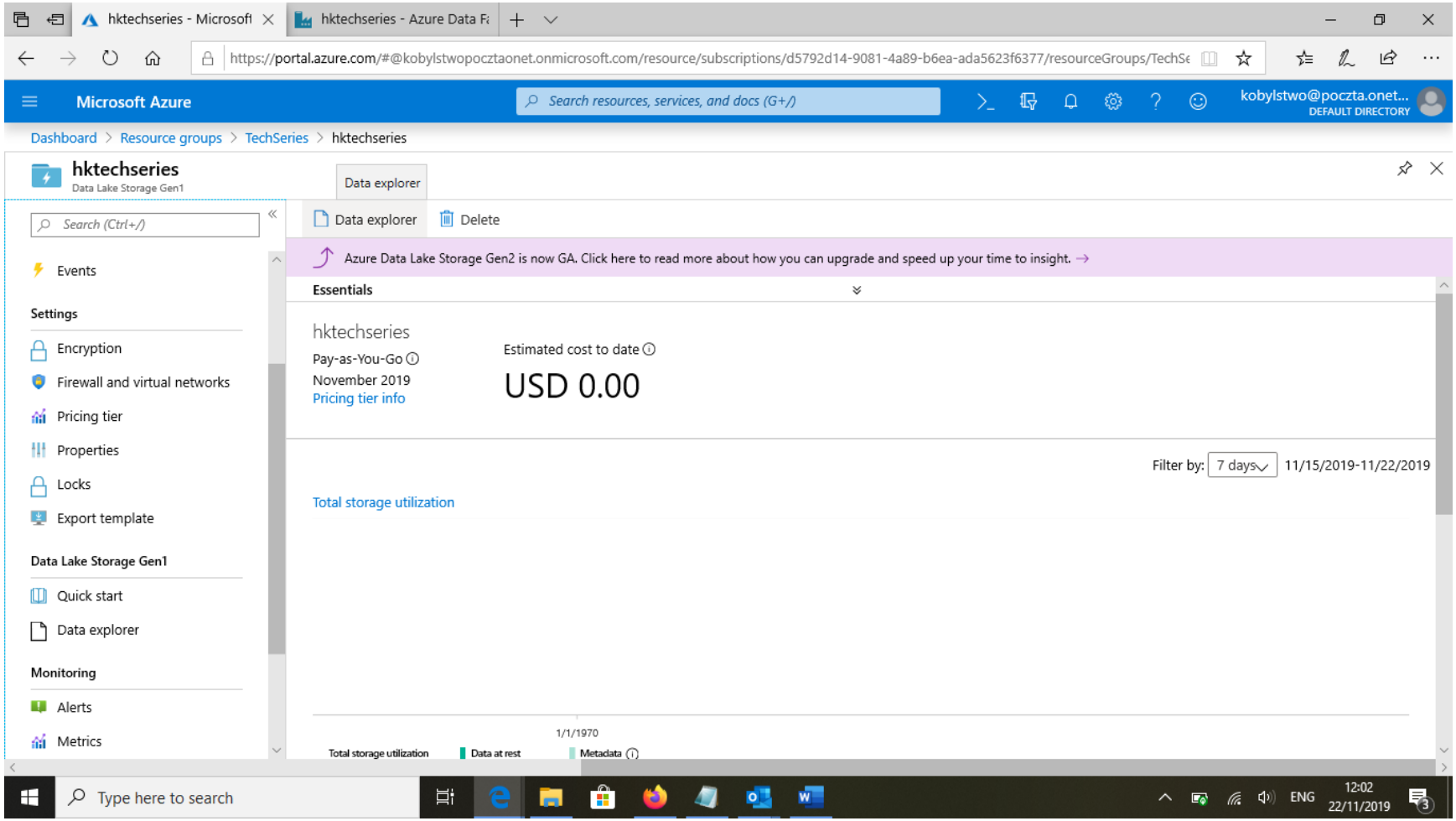


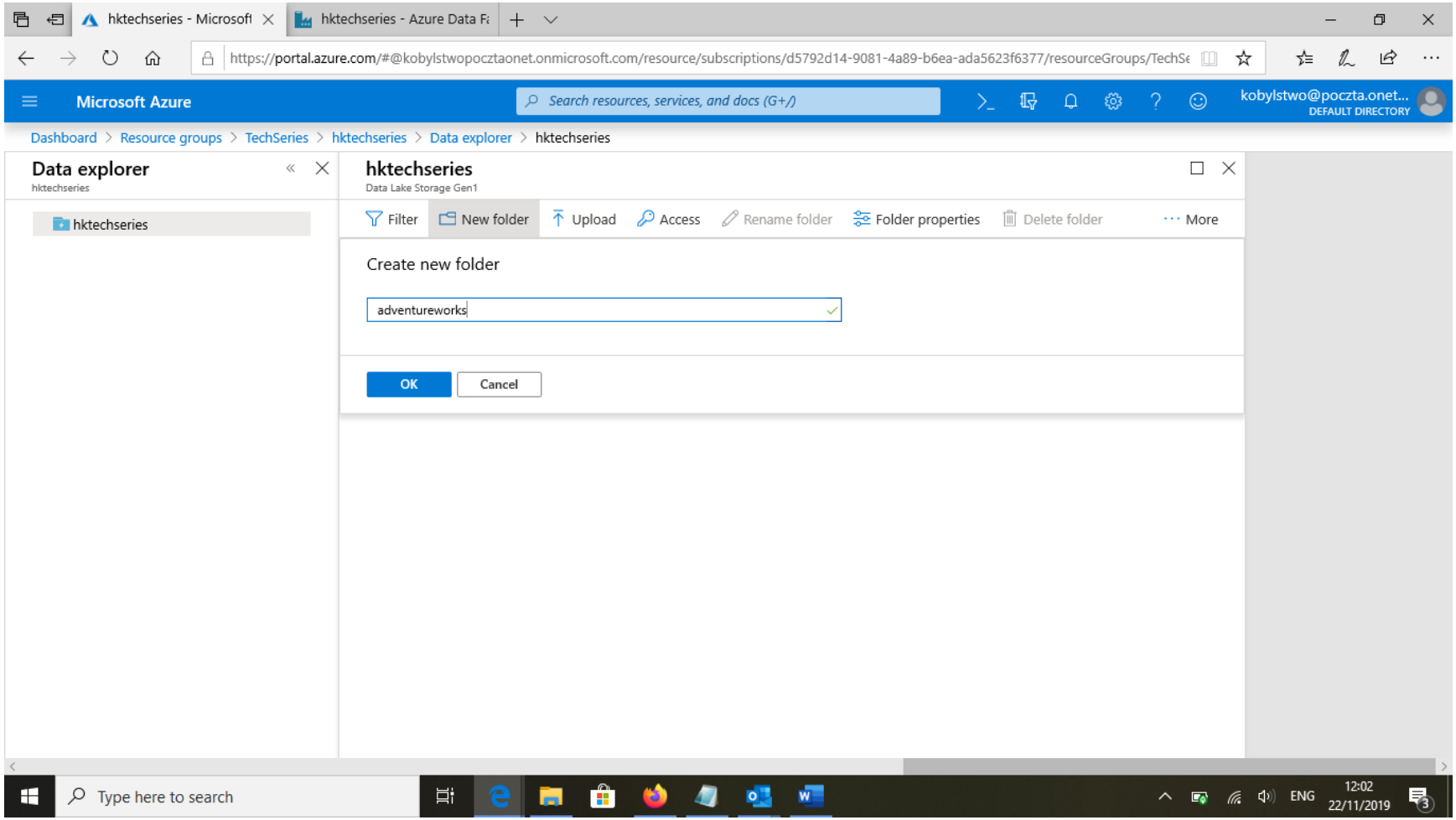
Don’t forget to disable GIT. We will not use it.



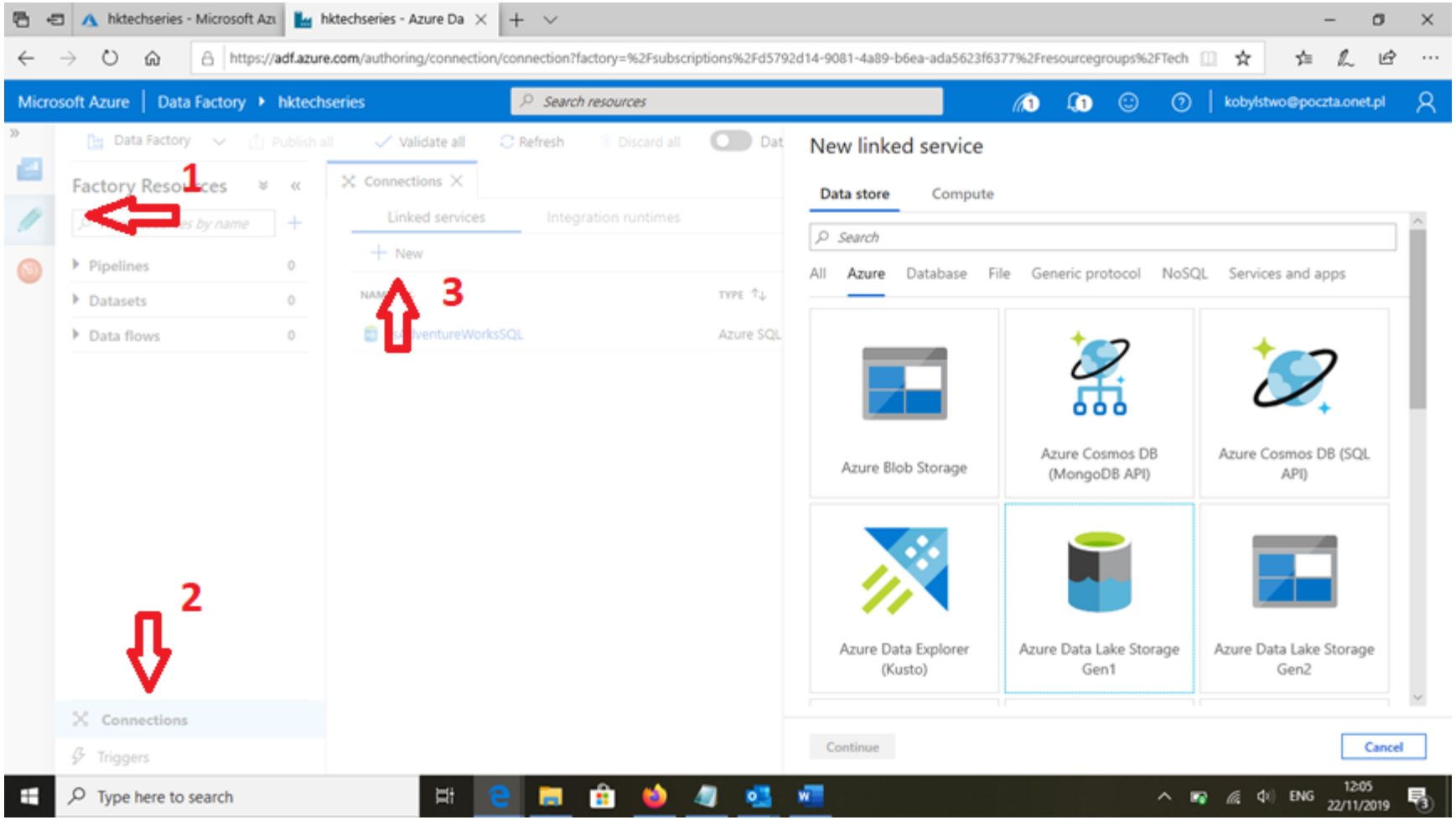
After it has been created you may check the interface by following the url: https://adf.azure.com/

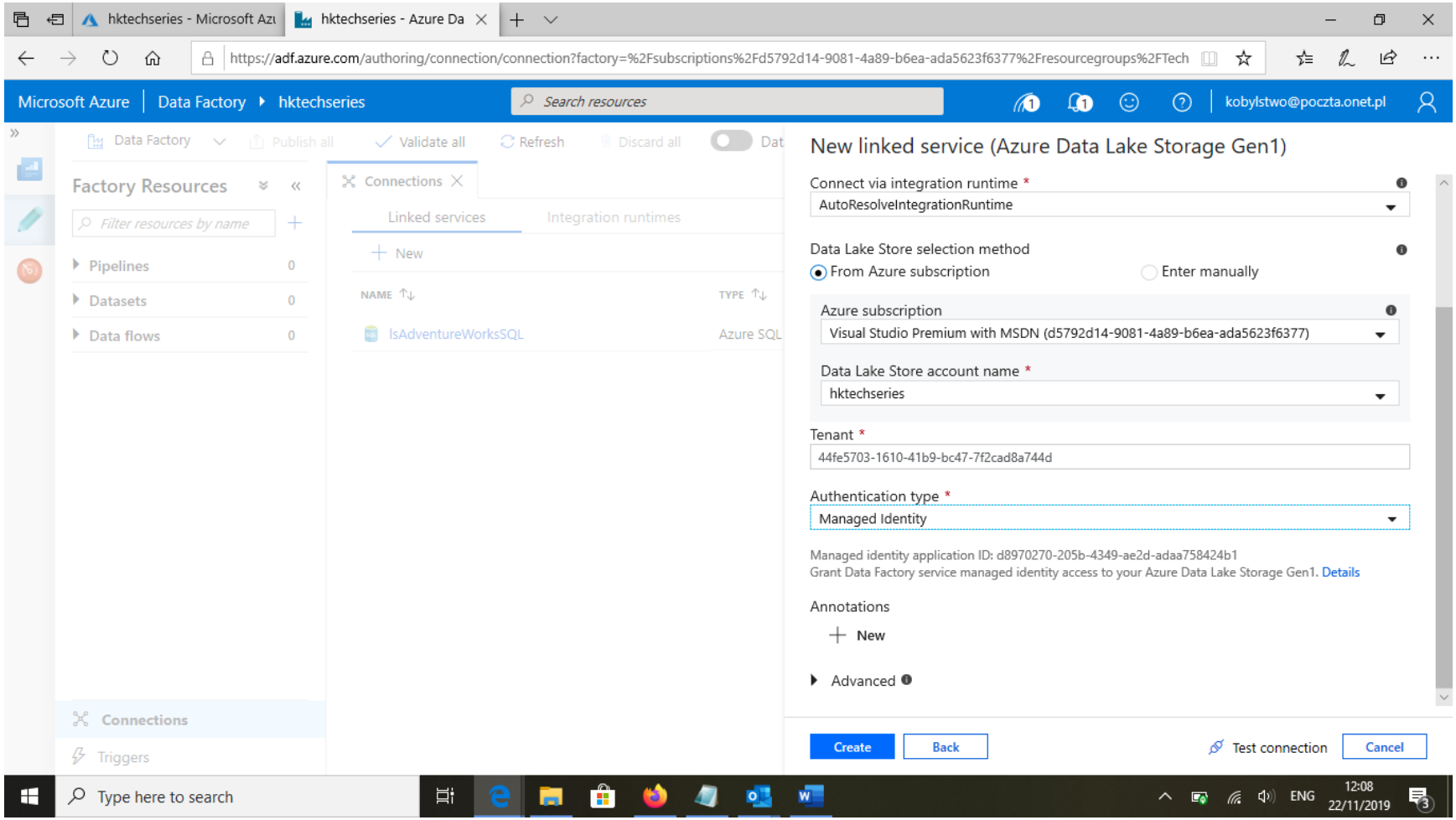
Now, create a folder in your instance of Data Lake. You may use “Data explorer” blade for that.




hktechseries - Microsoft
hktechseries - Azure Data Fa
https://portal.azure.com/#@kobylstwopocztaonet.onmicrosoft.com/resource/subscriptions/d5792d14-9081-4a89-b6ea-ada5623f6377/resourceGroups/TechSe
Microsoft Azure
Search resources, services, and docs (G+/)
kobylstwo@poczta.onet...
DEFAULT DIRECTORY
Dashboard > Resource groups > TechSeries > hktechseries > Data explorer > hktechseries
Data explorer
hktechseries
hktechseries
hktechseries
Data Lake Storage Gen1
Filter
New folder
Upload
Access
Rename folder
Folder properties
Delete folder
More
Create new folder
adventureworks
OK
Cancel
Type here to search
ENG
12:02
22/11/2019

Go to Data Factory. Create linked services in ADF for Data Lake (name it as lsDataLake) and Azure Synapse Analytics (name it as lsSynapse).



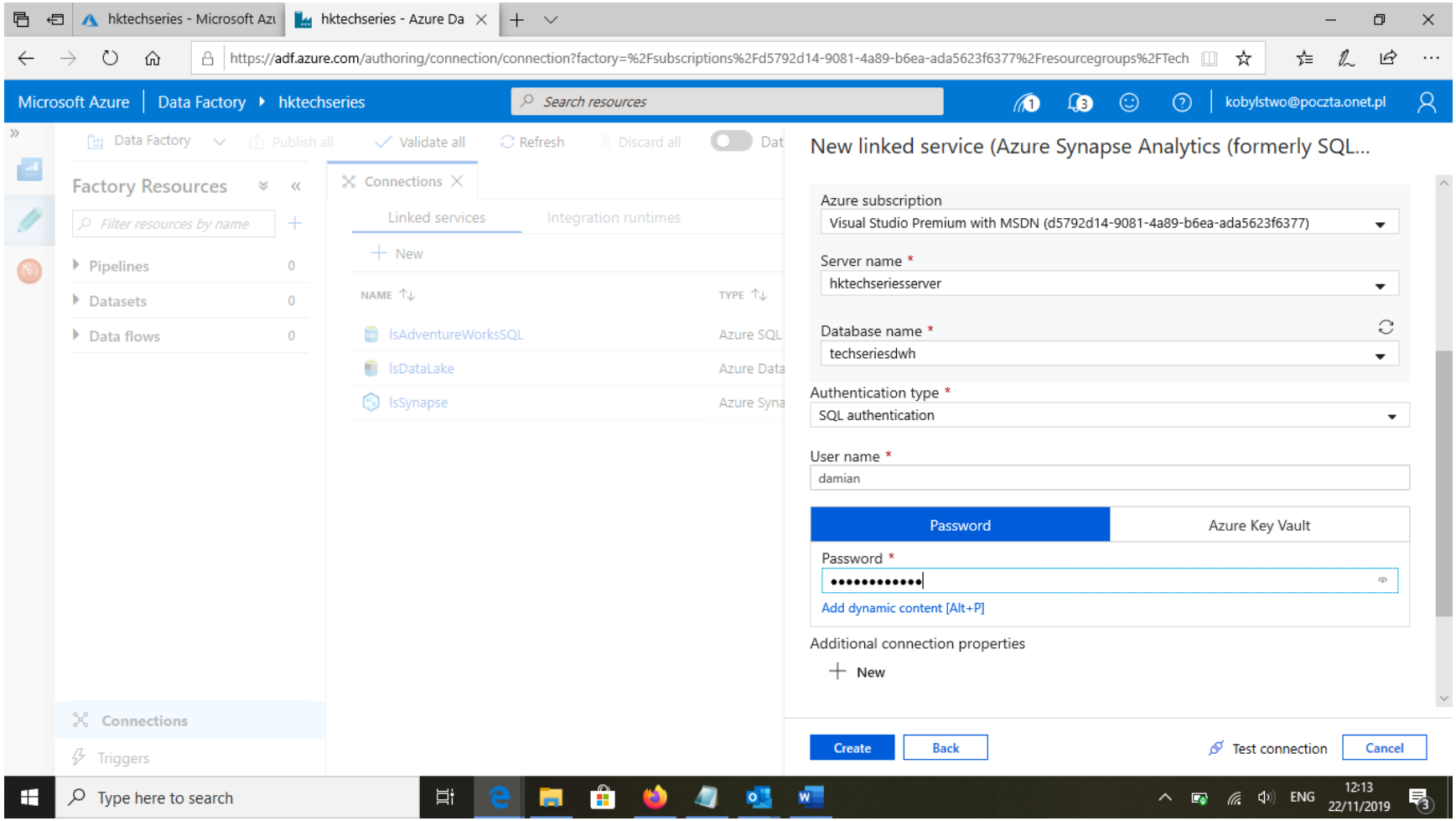


That object is defined by the following JSON code.

```
{
    "name": "lsDataLake",
    "type": "Microsoft.DataFactory/factories/linkedservices",
    "properties": {
        "annotations": [],
```

```
        "type": "AzureDataLakeStore",
        "typeProperties": {
            "dataLakeStoreUri": "https://hktechseries.azuredatalakestore.net/webhdfs/v1",
            "tenant": "44fe5703-1610-41b9-bc47-7f2cad8a744d",
            "subscriptionId": "d5792d14-9081-4a89-b6ea-ada5623f6377",
            "resourceGroupName": "TechSeries"
        }
    }
}
```

Now, a linked service for Synapse.



Go to your Azure SQL DB and create following views (use any client you want, Azure Data Studio, SQL Management Studio, etc.). It will create some analytics views on our OLTP database.

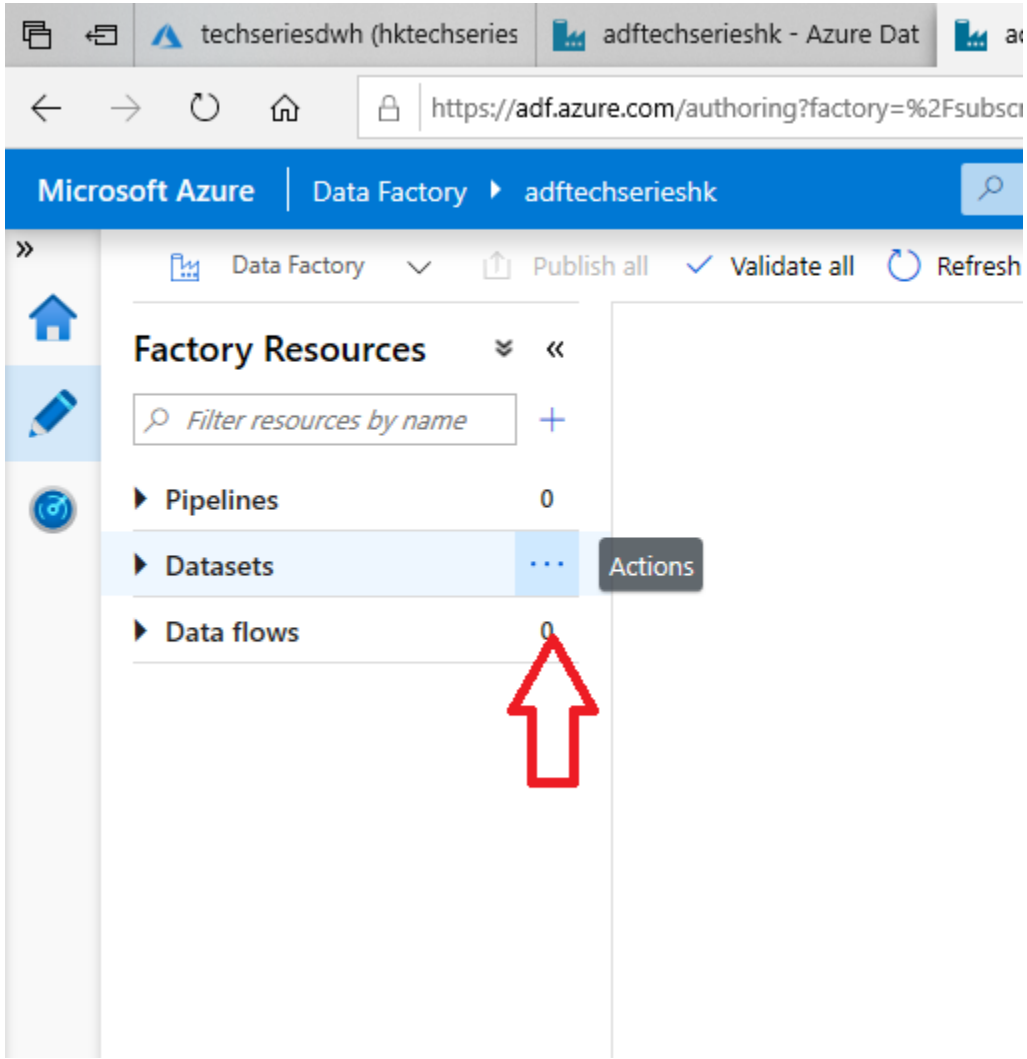
```
CREATE VIEW vSales
```

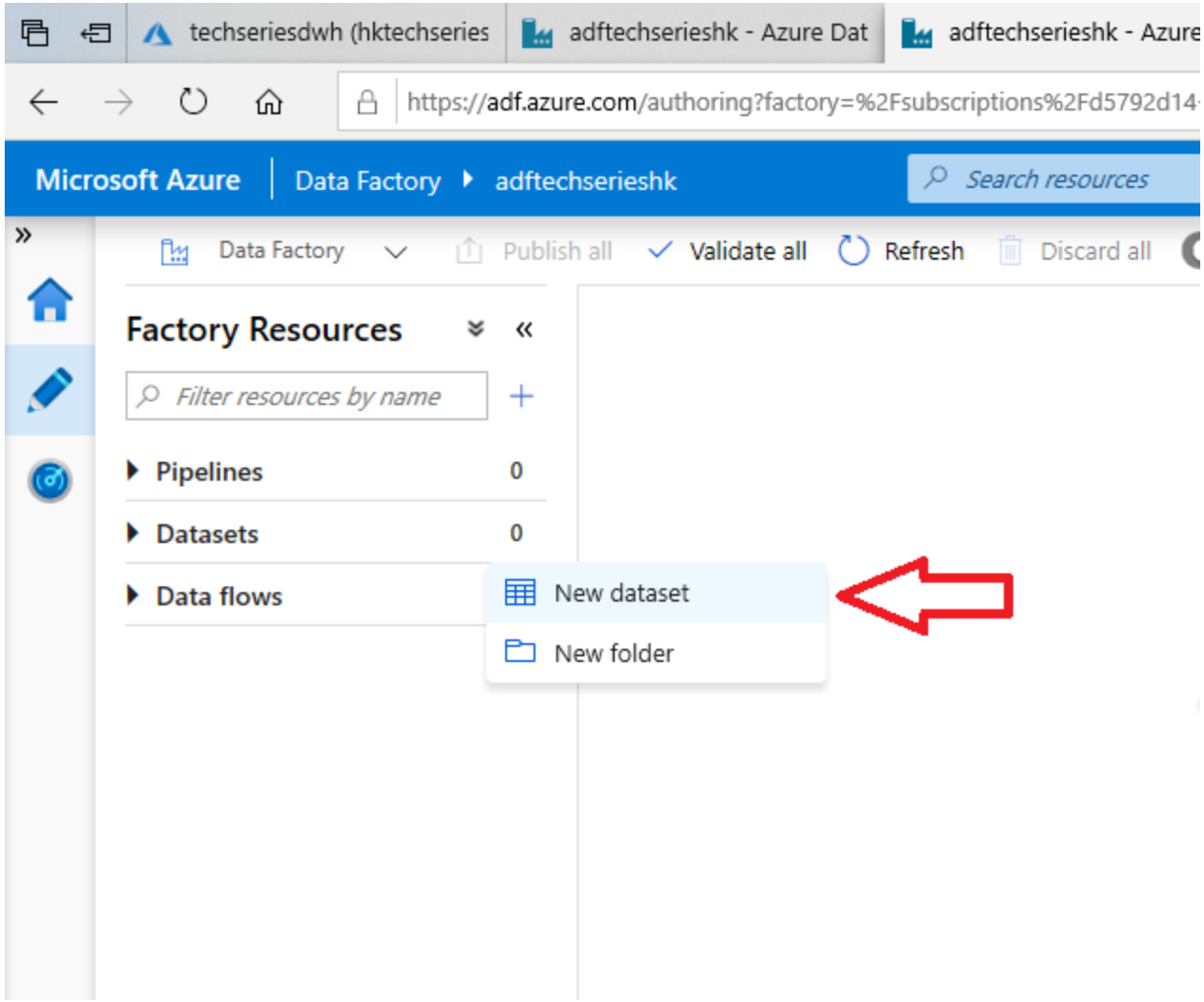
```
AS
SELECT oh.[OrderDate]
      ,oh.[SalesOrderNumber]
      ,oh.[CustomerID]
         ,od.ProductID
         ,od.OrderQty
         ,od.UnitPrice
  FROM [SalesLT].[SalesOrderHeader] oh
    INNER JOIN [SalesLT].[SalesOrderDetail] od
       ON oh.SalesOrderID = od.SalesOrderID;

CREATE VIEW vProduct
AS
SELECT p.[ProductID]
      ,p.[Name] AS Product
         ,pc.[Name] AS Category
      ,p.[ProductNumber]
      ,p.[Color]
      ,p.[StandardCost]
  FROM [SalesLT].[Product] p
  LEFT JOIN [SalesLT].[ProductCategory] pc
    ON p.ProductCategoryID = pc.ProductCategoryID;

CREATE VIEW vCustomer
AS
SELECT [CustomerID]
      ,[FirstName] + ' ' + [MiddleName] AS Customer
  FROM [SalesLT].[Customer];
```

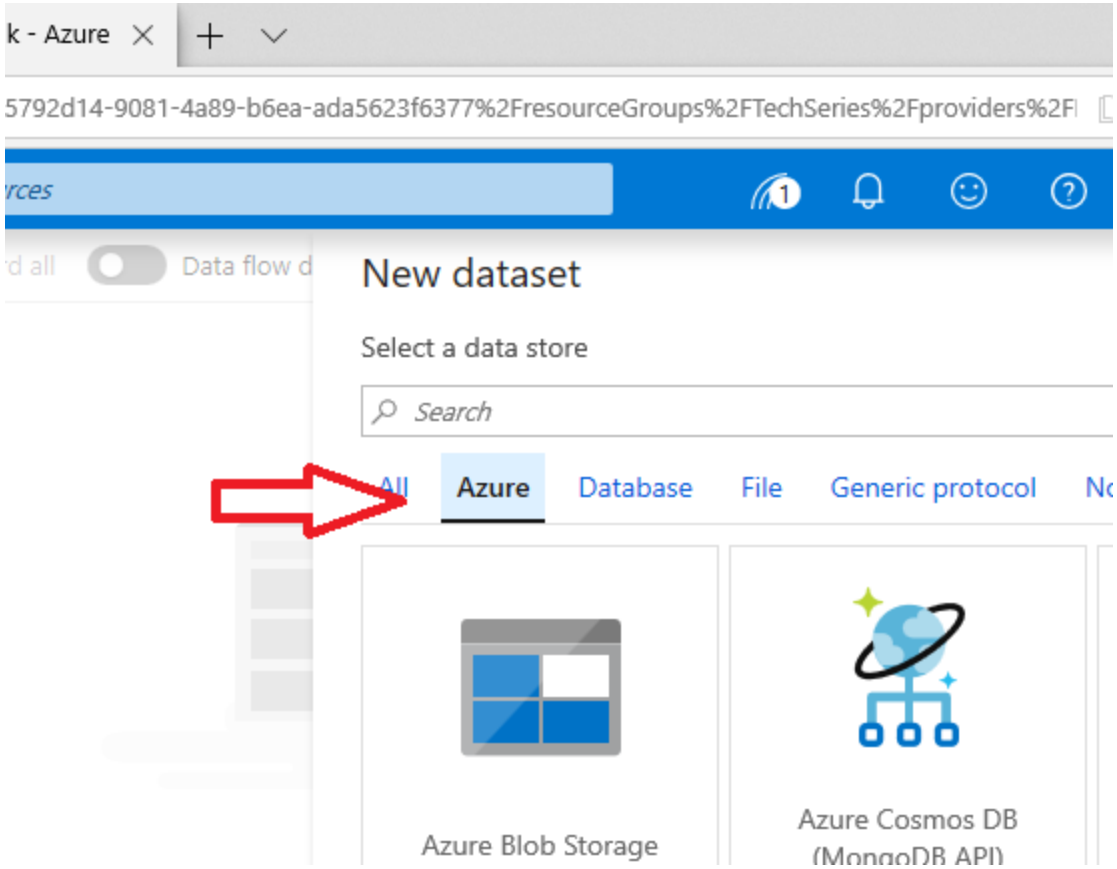
Go to Azure Data Factory, click on Author & Monitor. Alternatively you can go to https://adf.azure.com/

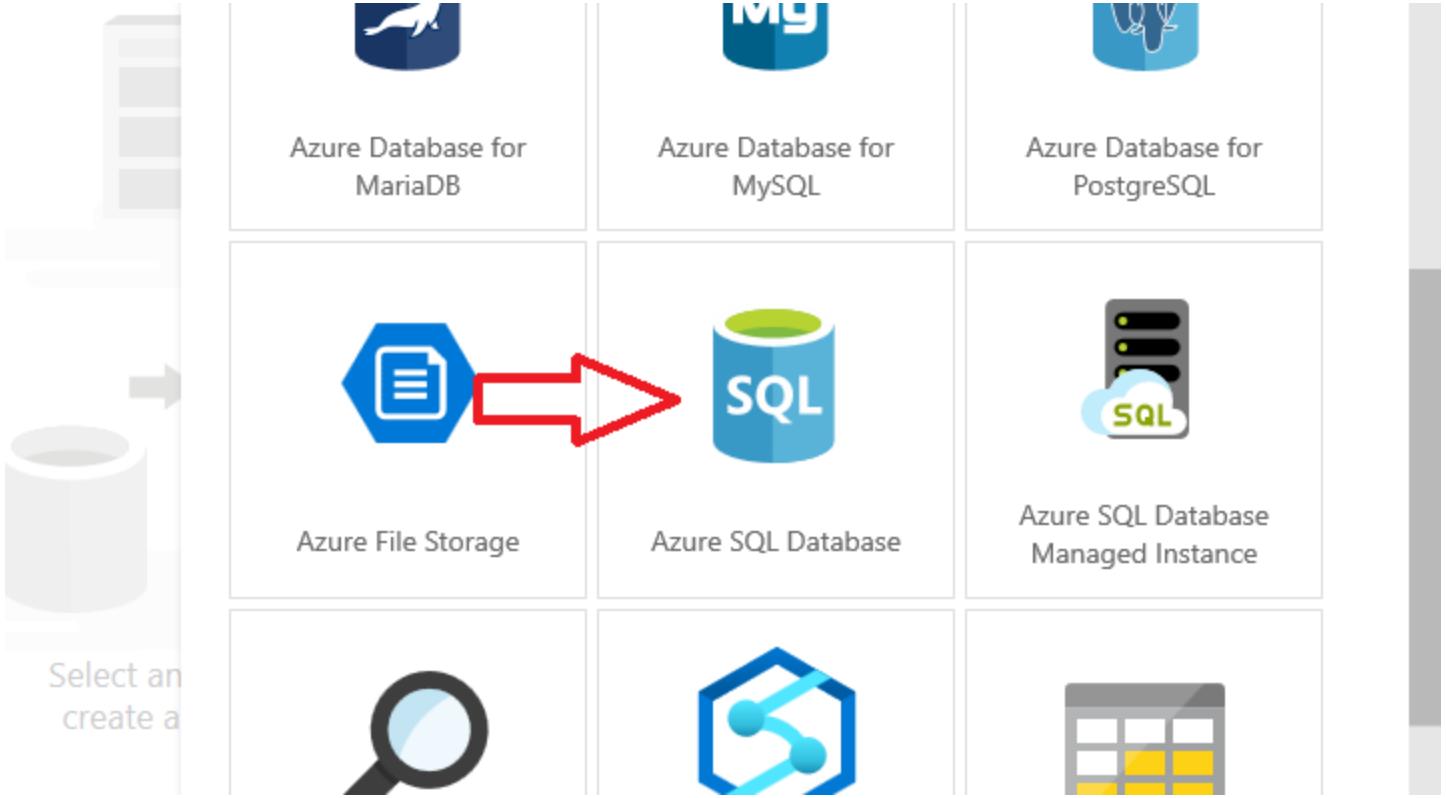
Create a connector (called here as Linked Service) to Azure SQL DB. We will be loading data from it. Just follow similar steps as for Data Lake and Synapse Analytics services.

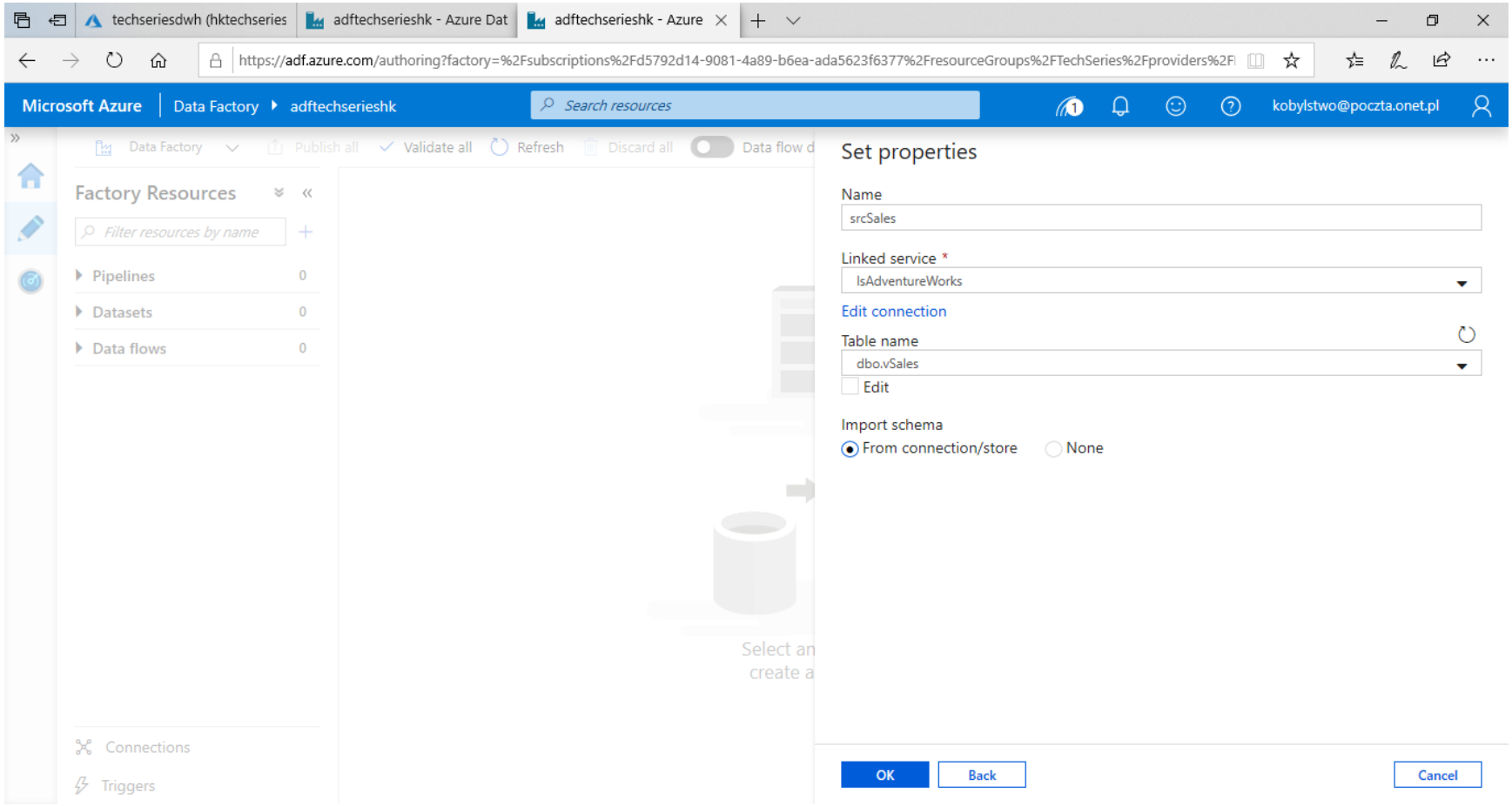
Create source datasets – for Azure SQL DB views created above.



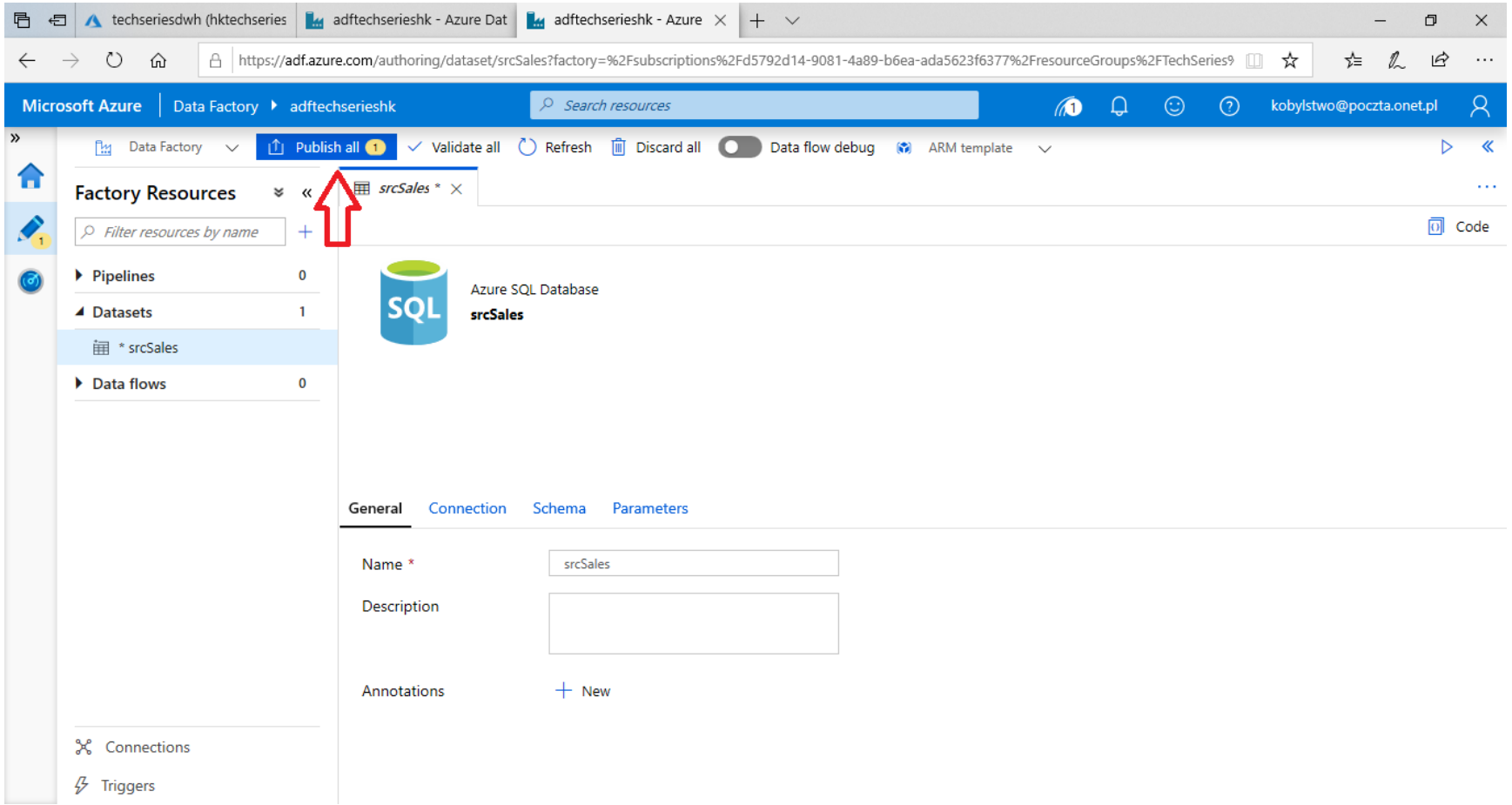

techseriesdwh (hktechseries
adftechserieshk - Azure Dat
adftechserieshk - Azure
https://adf.azure.com/authoring?factory=%2Fsubscriptions%2Fd5792d14
Microsoft Azure
Data Factory
adftechserieshk
Search resources
Data Factory
Publish all
Validate all
Refresh
Discard all
Factory Resources
Filter resources by name
Pipelines
0
Datasets
0
Data flows
New dataset
New folder


k - Azure
5792d14-9081-4a89-b6ea-ada5623f6377%2FresourceGroups%2FTechSeries%2Fproviders%2F
rces
1
d all
Data flow d
New dataset
Select a data store
Search
All
Azure
Database
File
Generic protocol
Azure Blob Storage
Azure Cosmos DB
(MongoDB API)

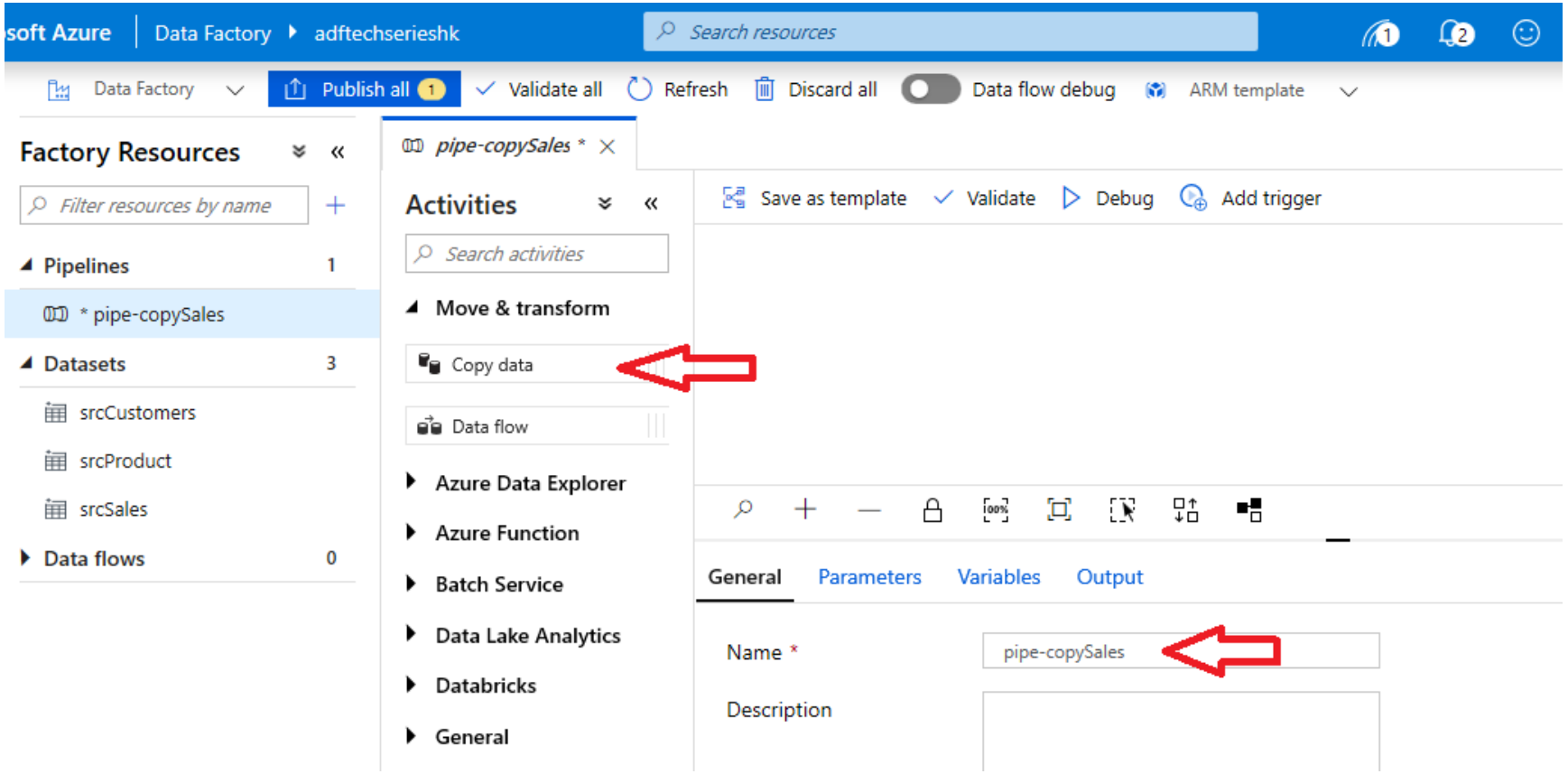

Azure Database for MariaDB
Azure Database for MySQL
Azure Database for PostgreSQL
Azure File Storage
Azure SQL Database
Azure SQL Database Managed Instance
Select an
create a


techseriesdwh (hktechseries
adftechserieshk - Azure Dat
adftechserieshk - Azure
https://adf.azure.com/authoring?factory=%2Fsubscriptions%2Fd5792d14-9081-4a89-b6ea-ada5623f6377%2FresourceGroups%2FTechSeries%2Fproviders%2F
Microsoft Azure
Data Factory
adftechserieshk
Search resources
kobylstwo@poczta.onet.pl
Data Factory
Publish all
Validate all
Refresh
Discard all
Data flow d
Factory Resources
Filter resources by name
Pipelines 0
Datasets 0
Data flows 0
Connections
Triggers
Select an
create a
Set properties
Name
srcSales
Linked service *
lsAdventureWorks
Edit connection
Table name
dbo.vSales
Edit
Import schema
From connection/store
None
OK
Back
Cancel

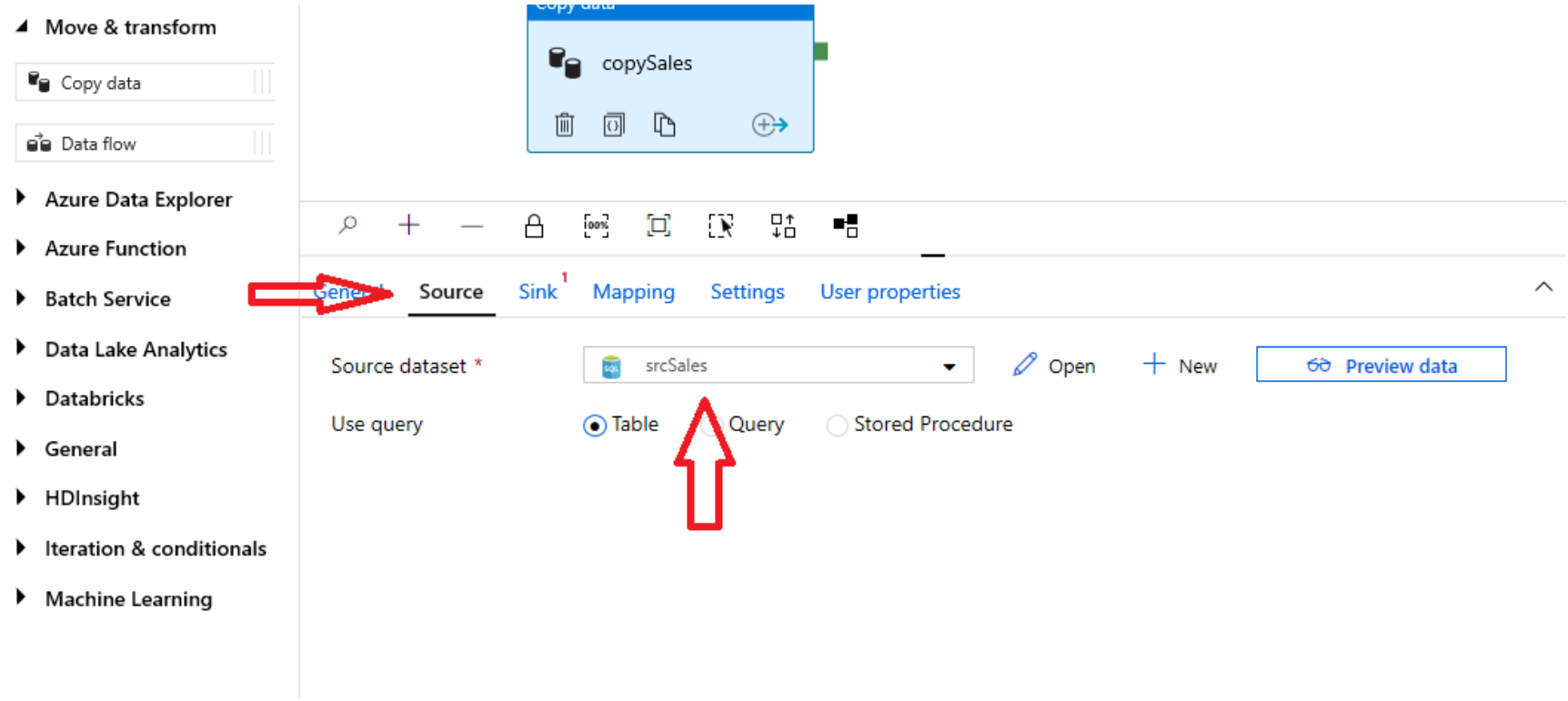
You don’t have to do anything else with it. Just publish it.



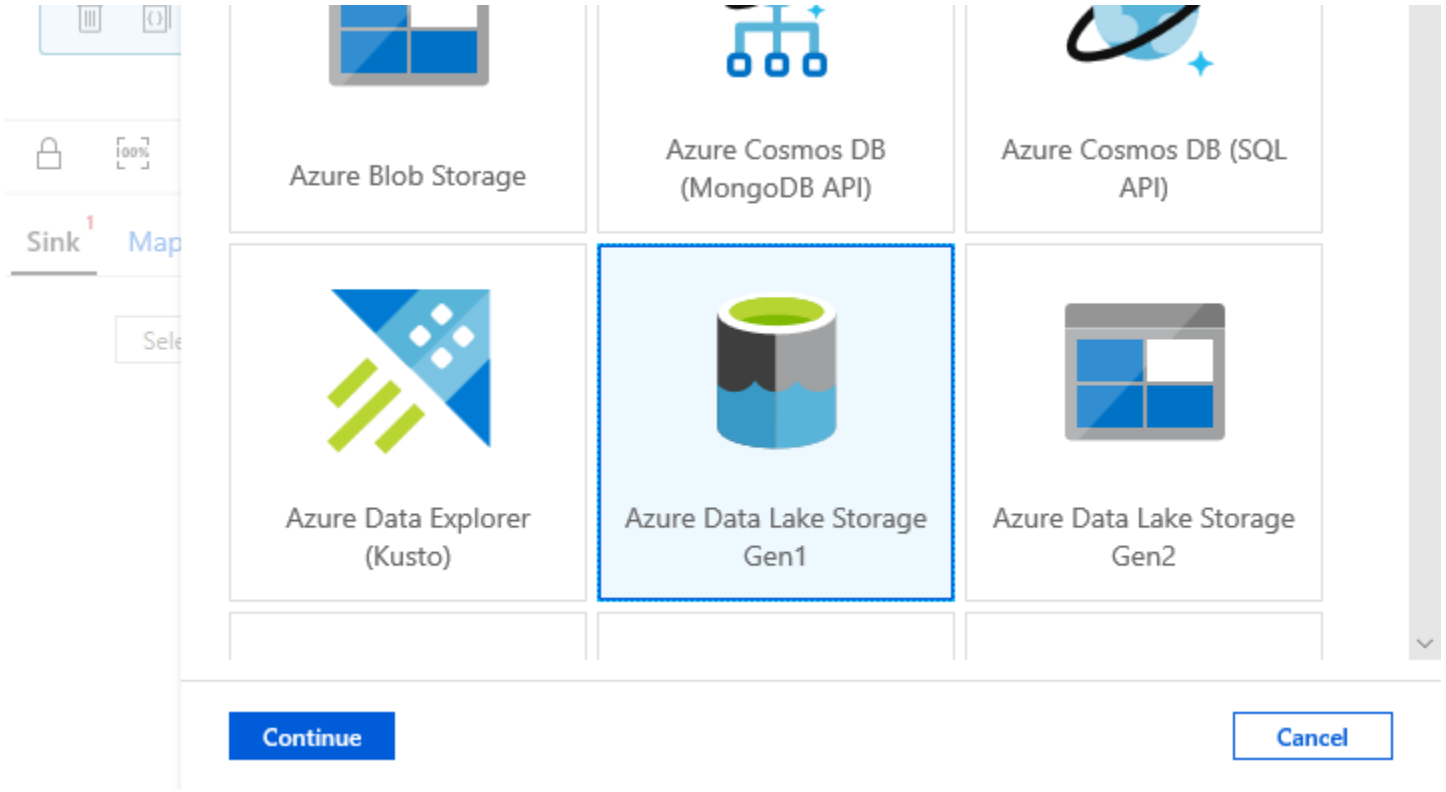
Now, create datasets for the other two views – Customer and Product. Once you have got them published create our first pipeline. Name it properly and drop the Copy data component on the canvas.

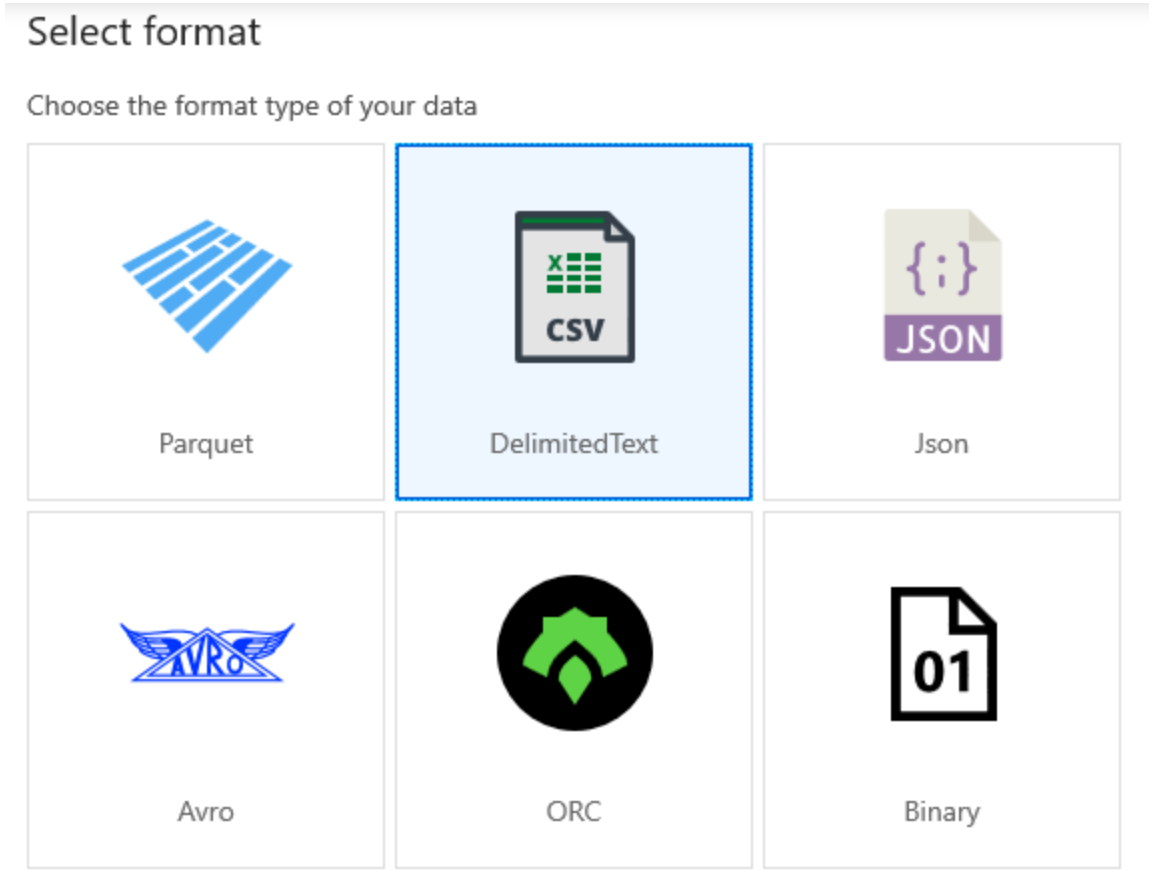


Rename the Copy data component, go to the Source tab and select the Azure SQL DB Sales dataset from the dropdown.

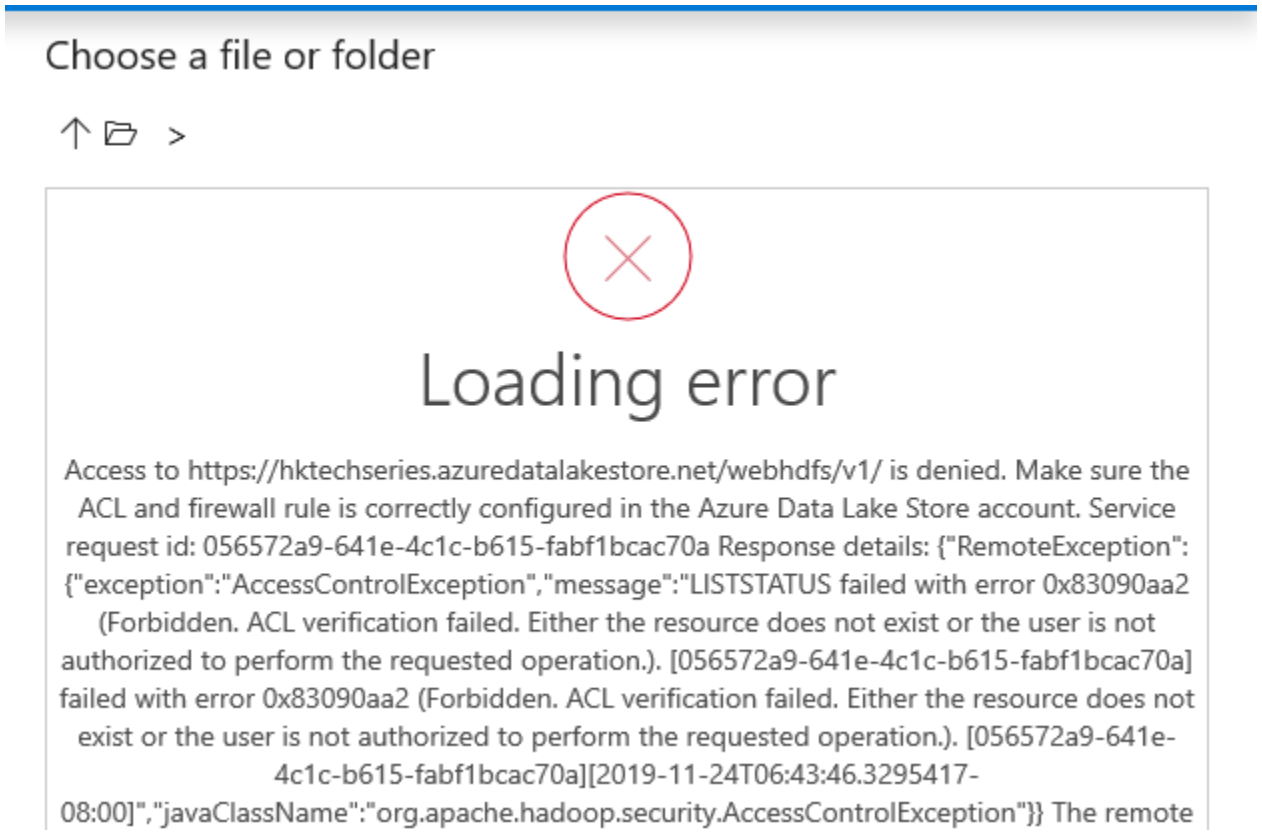


Next, go to the Sink tab and create a brand new Data Lake Gen1 dataset for Sales. Use CSV data format.



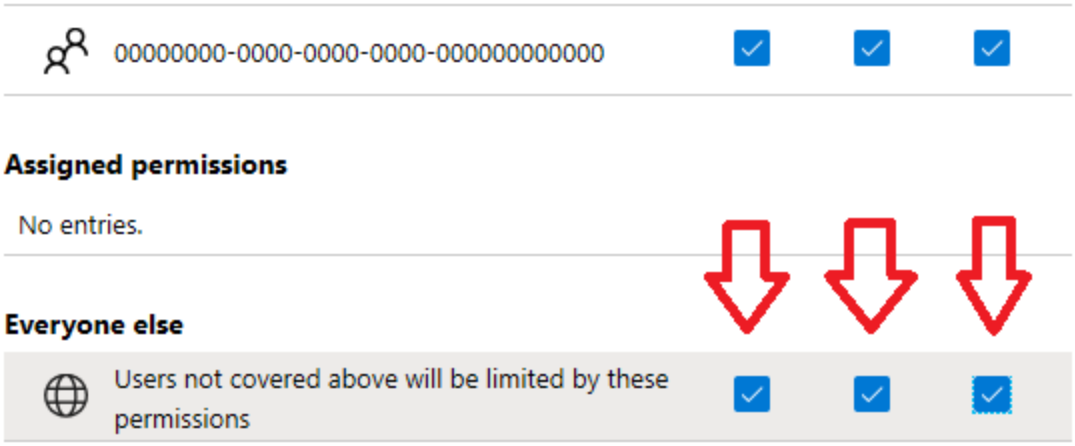

Select format
Choose the format type of your data
Parquet
CSV
DelimitedText
{;}
JSON
Json
AVRO
Avro
ORC
01
Binary

Name it as textSales, select your Data Lake linked service and hit Browse to select folder. You are going to see the access error.



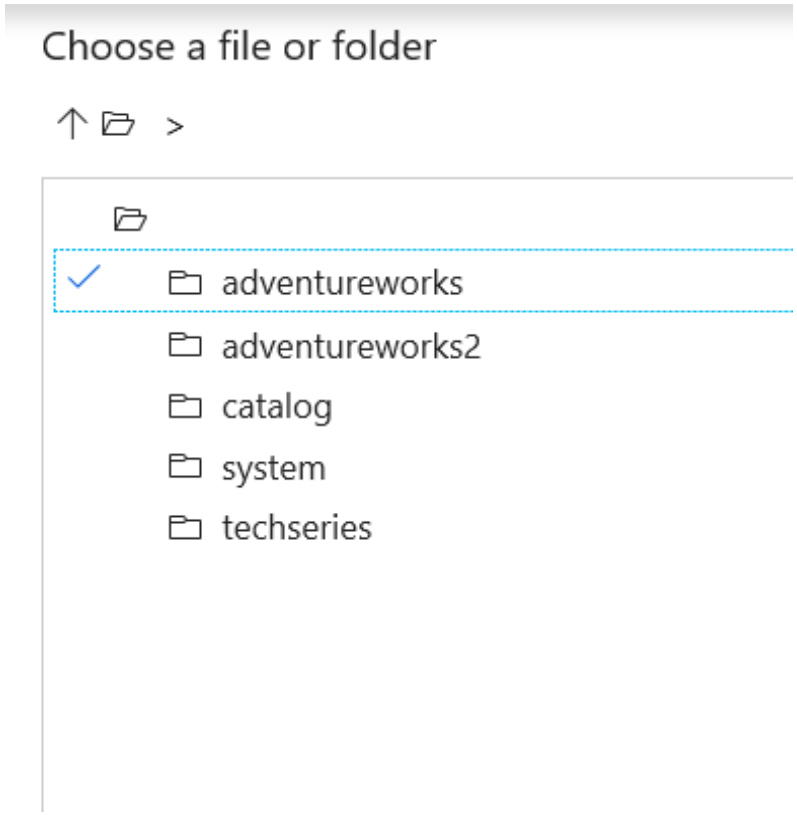
It means one thing – ADF has no rights to access ADLS resources. We are going to fix it the easy way in this tutorial but in your regular environment you should follow a secure path with stronger security bias.

Grant access to DL Gen1 folder. Got to the ADLS resource, use Data explorer and hit Access button in the root folder and tick those three boxes in the “Everyone else” section. Remember – this is not secure.

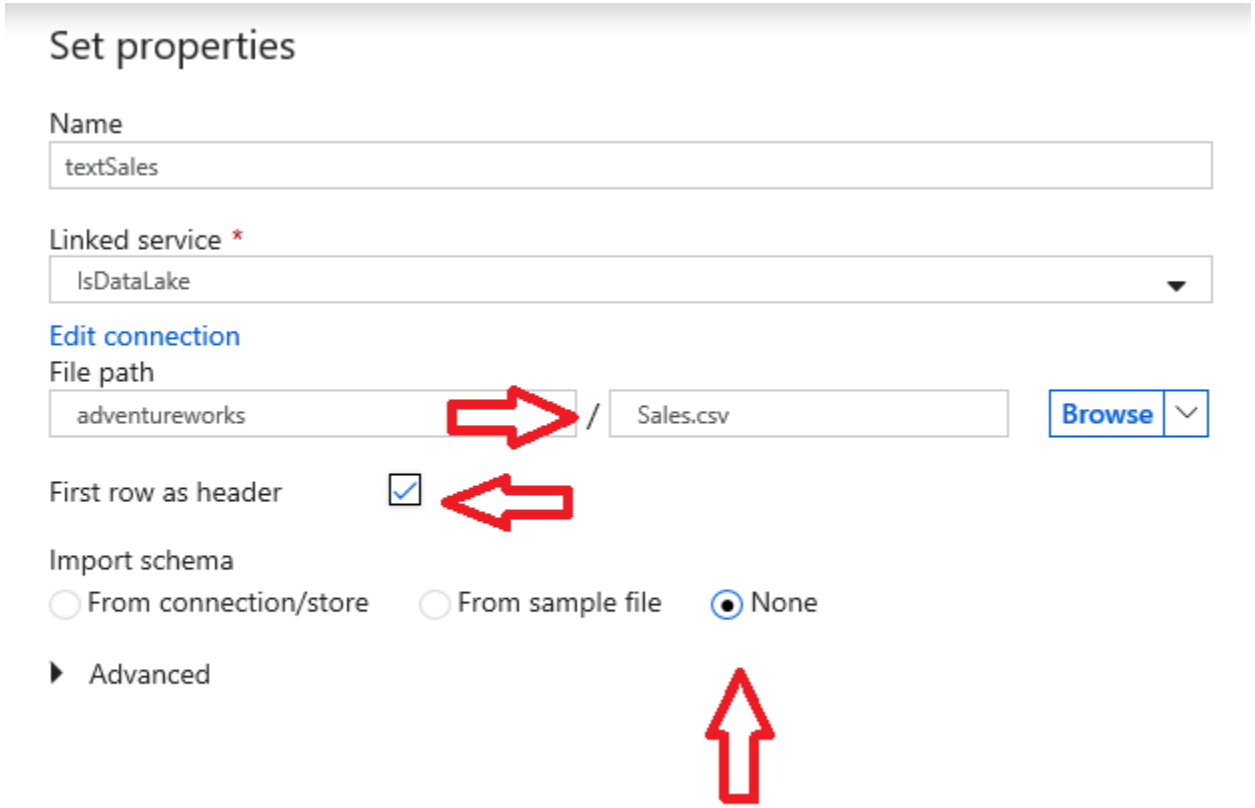


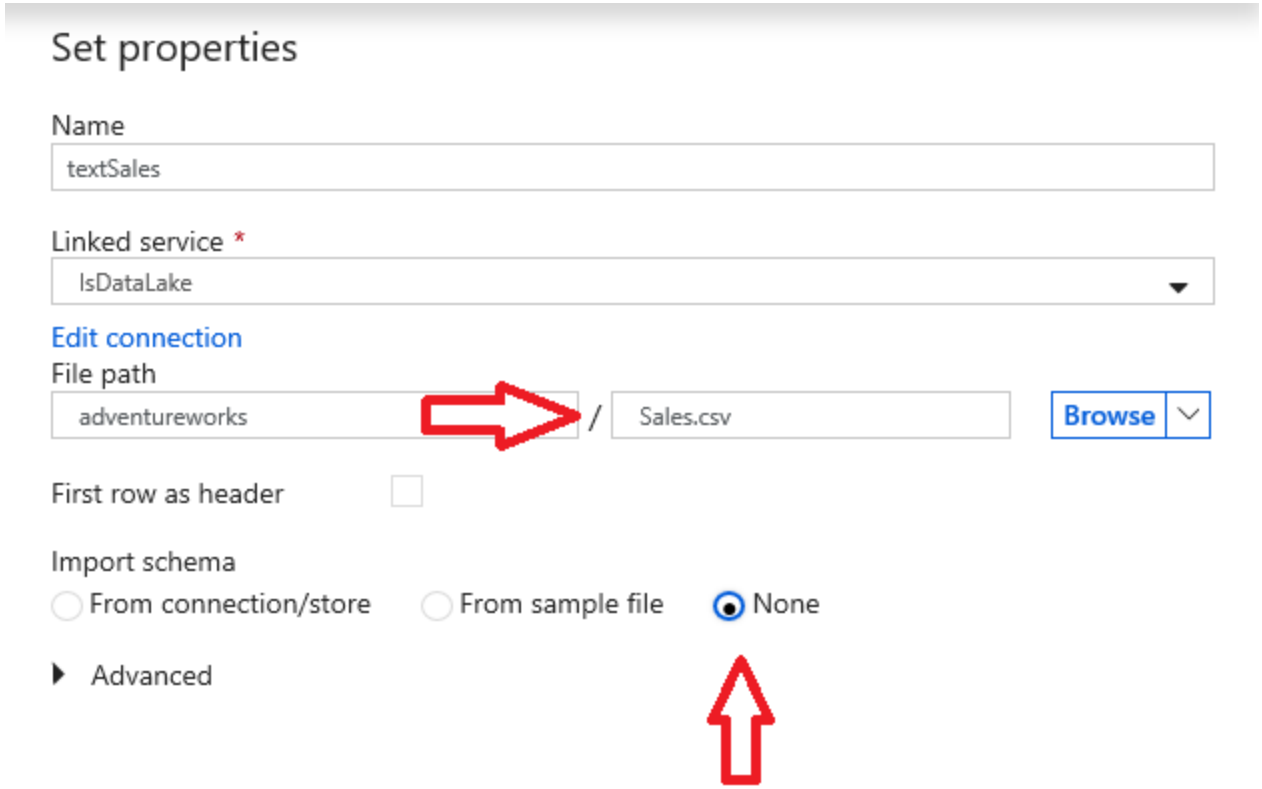
Hit “Save” button after ticking these boxes off.

Do the same thing for our ‘adventureworks” folder and go back to the ADF interface. Select the ‘adventureworks’ folder and press OK button.

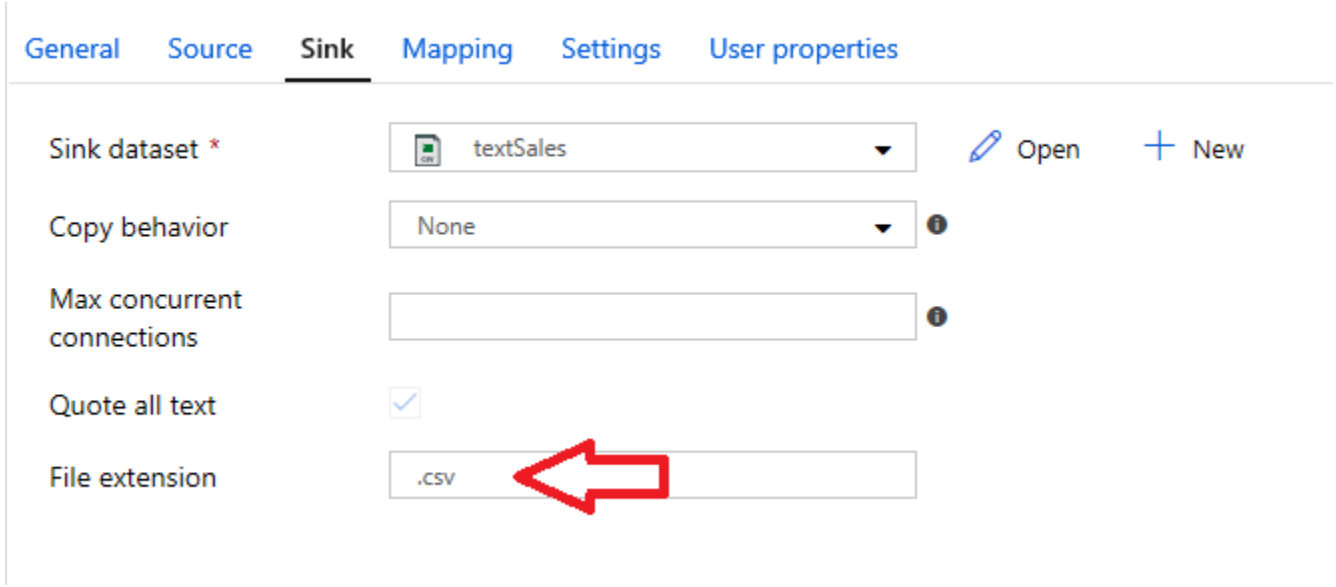

Choose a file or folder
>
adventureworks
adventureworks2
catalog
system
techseries

Enter ‘Sales.csv’ file name and select ‘None’ from the Import schema section as we do not have any structure for that file. Do not forget about setting first row as a column header.

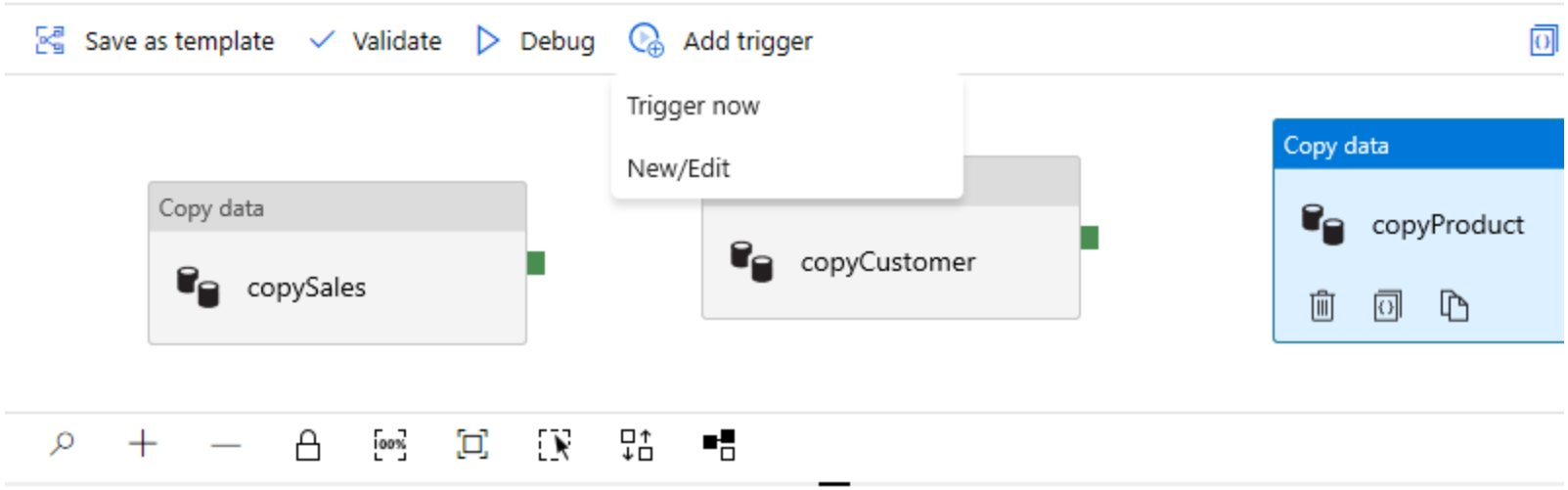



Set properties
Name
textSales
Linked service *
lsDataLake
Edit connection
File path
adventureworks
/
Sales.csv
Browse
First row as header
Import schema
From connection/store
From sample file
None
Advanced

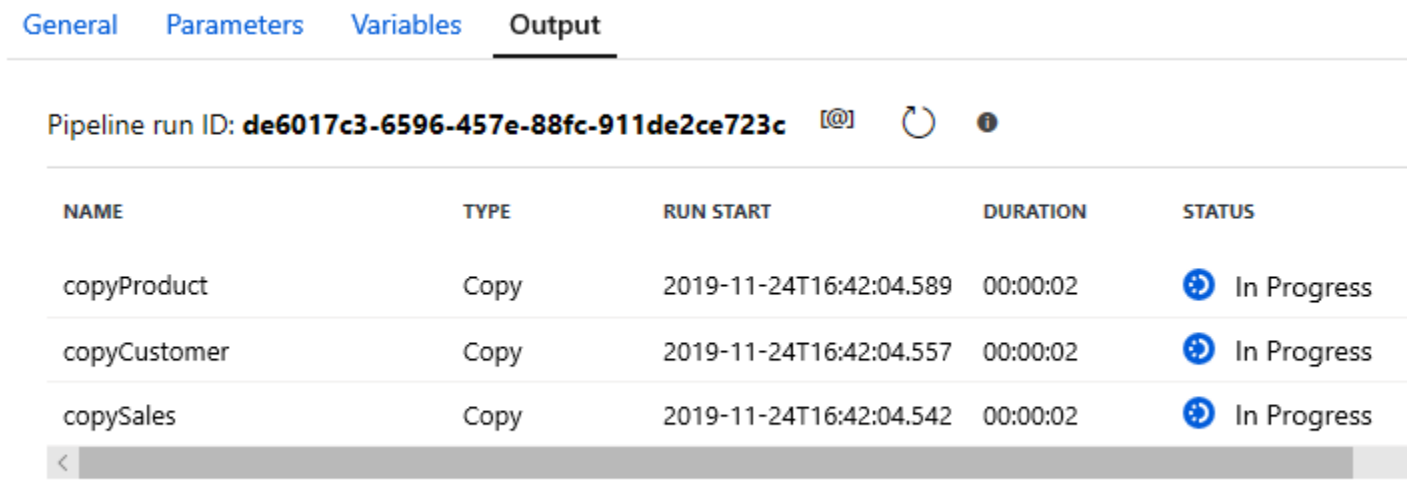
After creating your new sink dataset change File extension to '.csv' just to keep it consistent with the previous screen.



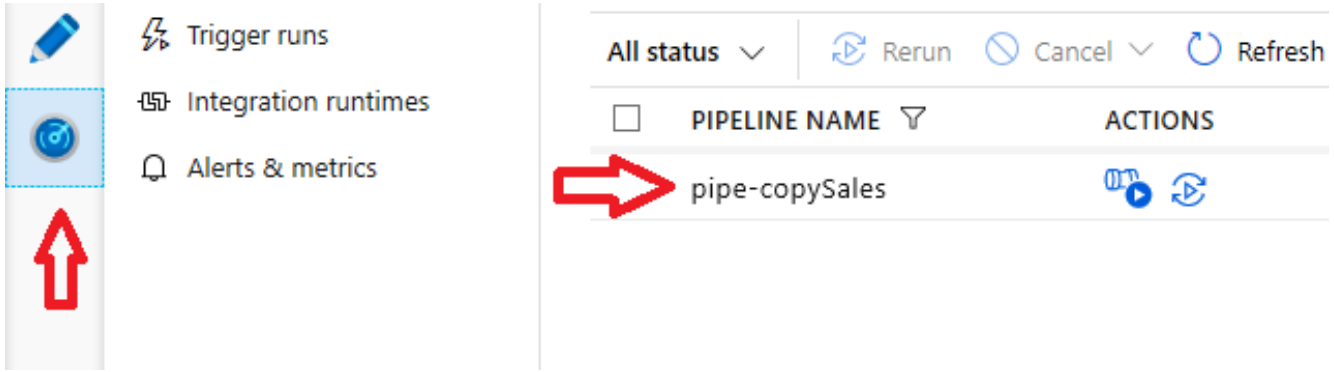
Your pipeline is ready and you may publish it. At the moment you need to add next two Copy data components to your pipeline. You do not have to set any mappings as your destination objects (files on Data Lake Storage instance are not existing as of this moment). Do not forget to publish your finished pipeline. To execute it you may either press the Debug button or Add trigger/Trigger now option. The Debug button enables testing your pipeline before it is published.



Your execution results will appear at the bottom of the screen.

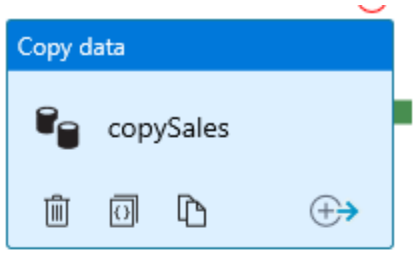


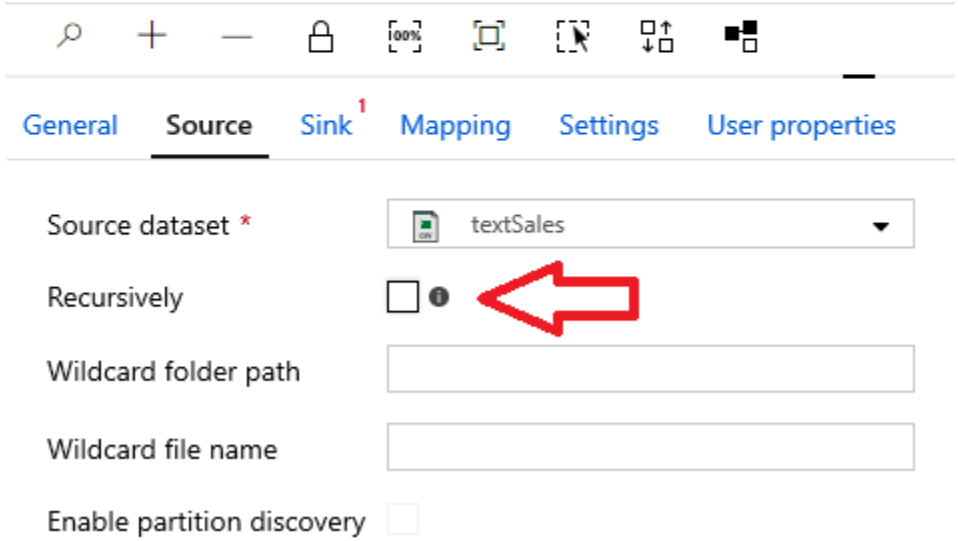
If you have executed the pipeline through a trigger you may observe its progress via the ADF monitoring section.



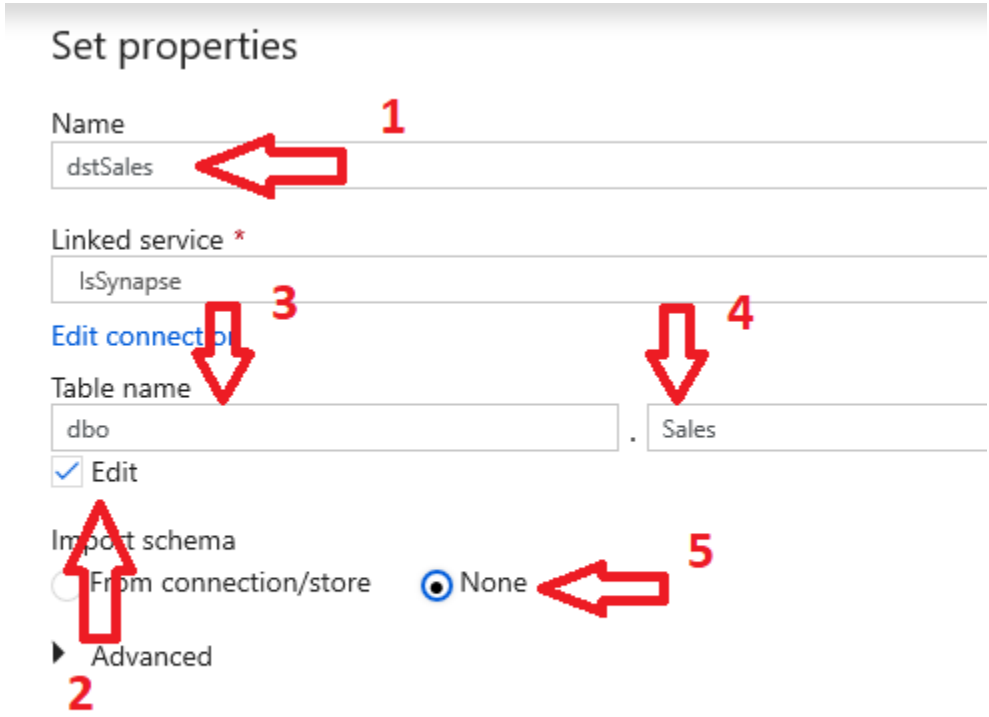
At this moment you should be ready to create a new pipeline which copies data from Data Lake to Synapse Analytics. Follow similar steps, remember that your source dataset names start with 'text*' and destination ones sit in the Synapse Analytics service. Following screens may be of some help here since your destination tables do not exist yet.

Do not forget to uncheck the ‘Recursively’ checkbox.

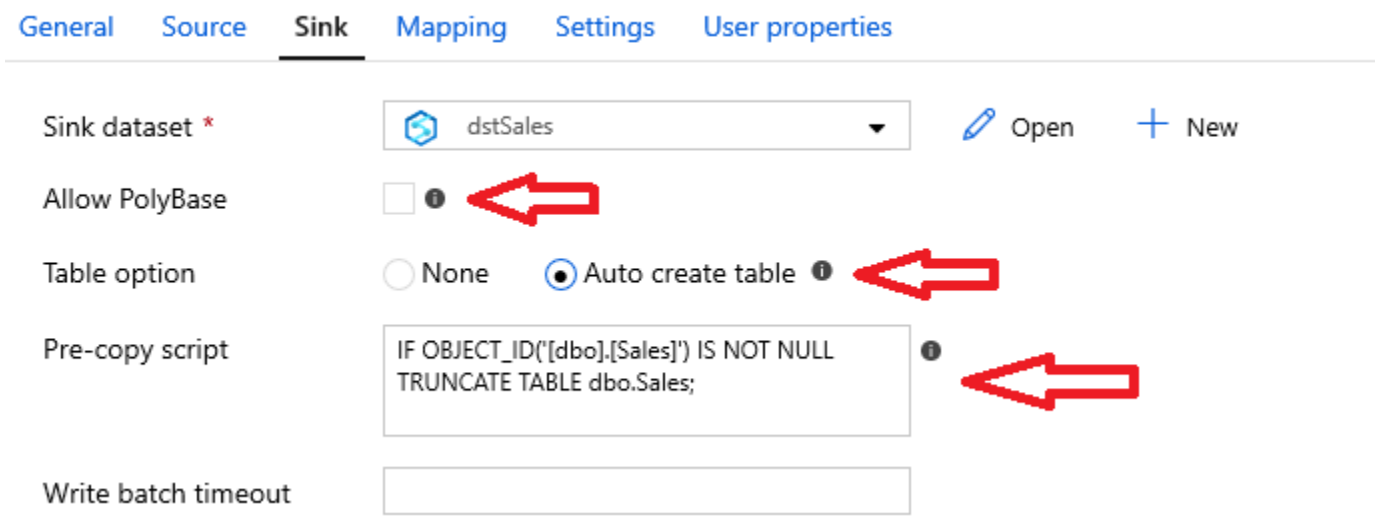




When setting properties of the new dataset pay attention to the following options:
- start its name with ‘dst’ (just to be consistent with the naming convention)
- check the ‘Edit’ box in order to enter non-existing table name
- put ‘dbo’ and ‘Sales’ in respective text boxes
- select ‘None’ option in the ‘Import schema’ section



Once the dataset is created you need to add few Sink options.



Follow similar steps for the other two tables (Customer and Product). Publish and trigger your new pipeline and you should be able to build your first report i.e. in Power BI.