

Azure Data Engineer Project

Table of Contents

Project overview	2
Key Objectives	2
Project Methodology	2
Project Benefits	3
Used Service/Tool in Azure:	4
Architecture of the project	5
Environment Setup	6
Resource group	6
What is a resource group in Azure?	6
How to create a resource group in Azure	6
Best practices for using resource groups	8
Resource group - service	8
Azure Data Factory (ADF)	8
Key Features of ADF	8
Benefits of Using ADF	9
Resource group - Project	9
Azure Databricks	11
Key Features of Azure Databricks	11
Benefits of Using Azure Databricks	11
Resource group - Project	12
Azure Key Vault	14
Why use Azure Key Vault?	14
Key Vault features	14
How to use Azure Key Vault	15
Resource group - Project	15
Storage account	18
Key features of Azure Storage Account:	18
Types of Azure Storage Account:	18
Benefits of using Azure Storage Account:	18
Resource group - Project	19
Data source - on-premise database	21
Data source: SMSS - SQL server	21
Create Account and user	27
Data Ingestion: Azure data factory - setup	27
Configuration run Integration Service	27
Configuration Linked services	30
Configuration Pipeline	32

Lookup	34
ForEach	36
Data Transformation: Azure Databricks - setup	42
Databricks in the ADF (azure data factory)	43
Configuration Linked service	43
Pipeline Configuration	46
PowerBI - load transformed data	51
Azure Blob storage	51
Azure Data lake storage gen2	53
Glossary	55
Azure Blob Storage	55
Azure Data Lake Storage Gen2	55
Azure Data Lake Storage Gen	55
Parquet	56
Blob	56
Avro	56
Delta Format	56

Project overview

This data engineering project aims to migrate a company's on-premises database to Azure, leveraging Azure Data Factory for data ingestion, transformation, and storage. The project will implement a three-stage storage strategy, consisting of bronze, silver, and gold data layers (Medalion architecture). Bronze data will represent raw data extracted from the source database, silver data will undergo data cleansing, transformation, and enrichment, while gold data will serve as the aggregated and standardized data source for Power BI analytics. Azure Databricks will be employed for data transformation tasks.

Key Objectives

1. Migrate data from on-premises database to Azure: Utilize Azure Data Factory to seamlessly transfer data from the on-premises database to Azure storage accounts.
2. Implement a three-stage data storage strategy: Establish a bronze, silver, and gold data layer to handle raw, transformed, and aggregated data, respectively.
3. Leverage Azure Databricks for data transformation: Employ Azure Databricks' Apache Spark engine to perform data cleansing, transformation, and enrichment tasks.
4. Prepare data for Power BI analytics: Ensure that the gold data layer is in a format suitable for loading into Power BI dashboards and reports.

Project Methodology

The project will follow a structured methodology, encompassing the following phases:

1. Data Assessment and Planning: Thoroughly assess the existing on-premises database, identifying data sources, data volumes, and data quality issues. Plan the data migration process, including data storage locations and transformation strategies.
2. Data Ingestion: Utilize Azure Data Factory to establish data pipelines for extracting data from the on-premises database, loading it into Azure storage accounts, and ensuring data consistency and integrity.
3. Data Transformation: Employ Azure Databricks to perform data cleansing, transformation, and enrichment tasks within the silver data layer. This may involve data cleansing, data format conversion, data enrichment with external data sources, and data validation.

4. **Data Storage and Aggregation:** Move transformed data from the silver layer to the gold layer, which will serve as the centralized data source for Power BI analytics. Aggregate data to improve performance and reduce data volume.
5. **Data Quality Monitoring:** Establish data quality monitoring processes to ensure the accuracy, consistency, and completeness of data throughout the data lifecycle. Implement alerts and notifications to identify and address data quality issues promptly.
6. **Power BI Integration:** Prepare the gold data layer for loading into Power BI dashboards and reports. Ensure data alignment with Power BI data models and visualizations.
7. **Data Governance and Security:** Implement data governance policies to control access to sensitive data, enforce data quality standards, and adhere to regulatory compliance requirements. Utilize Azure Key Vault to securely store and manage data access credentials.

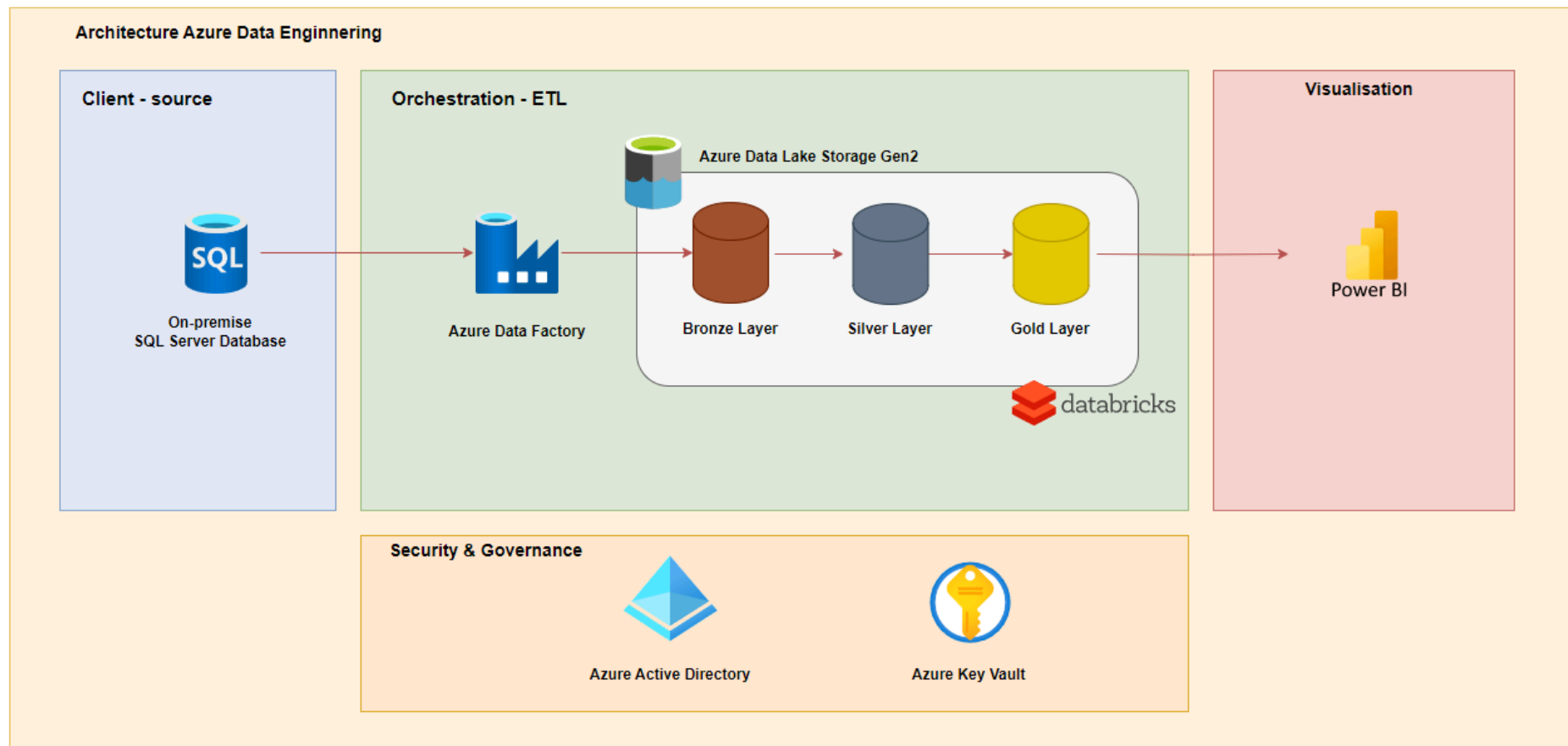
Project Benefits

1. **Reduced Operational Costs:** By migrating data to Azure, the company can eliminate the need for on-premises infrastructure and associated maintenance costs.
2. **Enhanced Scalability:** Azure provides the flexibility to scale data storage and processing capabilities to meet changing business needs.
3. **Improved Data Availability:** Azure's global infrastructure ensures high availability and data durability, minimizing the risk of data loss or downtime.
4. **Enhanced Data Quality:** The three-stage data storage approach facilitates data cleansing, transformation, and aggregation, improving data quality for downstream analysis.
5. **Accelerated Analytics:** Power BI provides a powerful and user-friendly platform for analyzing and visualizing data, enabling informed decision-making.

Used Service/Tool in Azure:

- Azure data factory
- Storage Account
- Azure Databricks
- Azure key vault
- Power BI

Architecture of the project



Environment Setup

Resource group

What is a resource group in Azure?

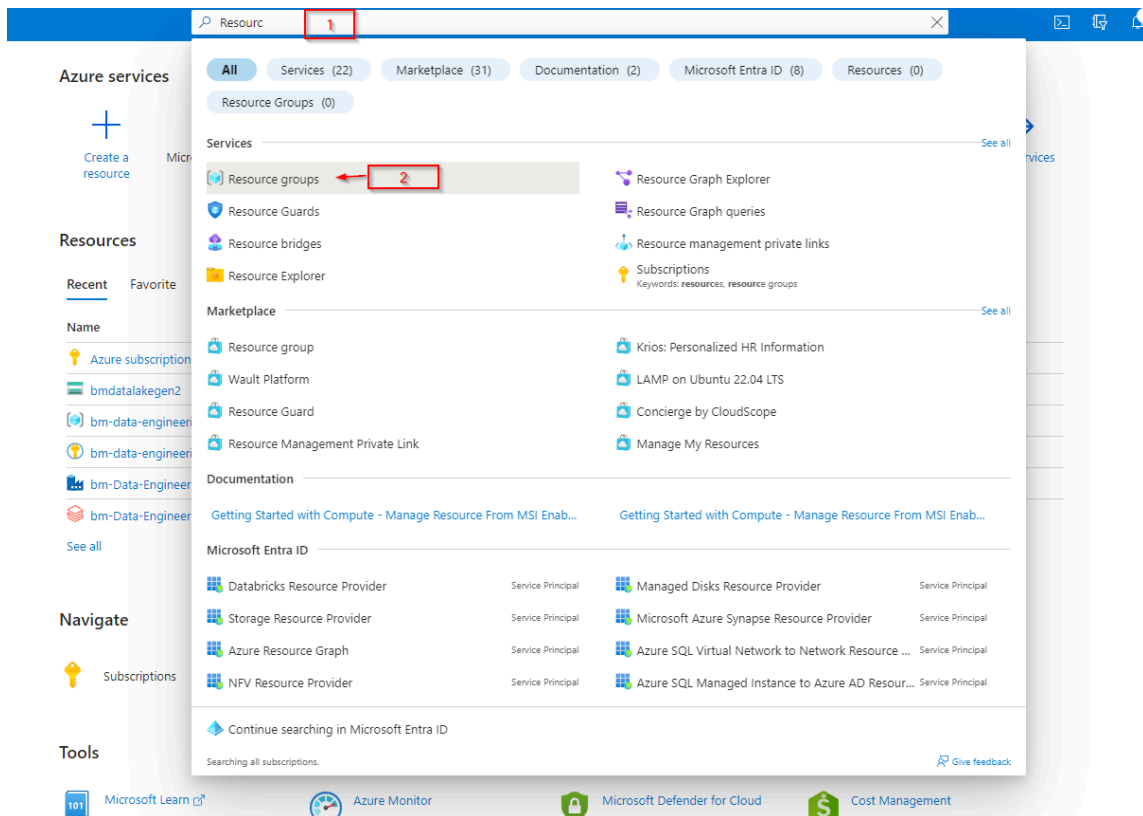
A resource group is a logical container that groups related Azure resources together. This includes resources such as virtual machines (VMs), storage accounts, databases, web apps, and more. Resource groups are essential for organizing and managing your Azure infrastructure, and they offer several benefits, including:

- **Simplified management:** Resource groups allow you to manage and monitor your resources as a single unit, which can make it easier to keep track of your cloud resources and ensure that they are all working properly.
- **Improved security:** You can control access to resources at the resource group level, which helps to protect your data and applications from unauthorized access.
- **Cost control:** Resource groups can help you to track your Azure spending by providing you with a consolidated view of your resource costs.
- **Consistent deployment:** Resource groups can be used to automate the deployment of Azure resources, which can help to ensure that your deployments are consistent and repeatable.

How to create a resource group in Azure

Creating a resource group in Azure is a simple process that can be done from the Azure portal. Here are the steps on how to create a resource group:

1. Navigate to the Azure portal and sign in to your account.
2. Type in "search bar" on the top "Resource group".
3. Click on the "Resource groups" icon in the dropdown menu.
4. Click on the "Create" button.
5. Enter a name for your resource group and a location for your resources.
6. Click on the "Review + create" button.
7. Review the details of your resource group and click on the "Create" button to create the resource group.



Create a resource group

Basics Tags Review + create

Resource group - A container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group. You decide how you want to allocate resources to resource groups based on what makes the most sense for your organization. [Learn more](#)

Project details

Subscription * 3

Resource group * 4

Resource details

Region * 5

6

Best practices for using resource groups

Here are some best practices for using resource groups in Azure:

- Create a separate resource group for each application or environment. This will help you to keep your resources organized and make it easier to manage your cloud infrastructure.
- Use a consistent naming convention for your resource groups. This will make it easier to find and identify your resources.
- Assign appropriate permissions to users and groups. This will help to protect your resources from unauthorized access.
- Monitor your resource groups regularly. This will help you to identify and resolve any issues with your resources.

Resource group - service

Azure Data Factory (ADF)

Azure Data Factory (ADF) is a cloud-based data integration service provided by Microsoft for building, managing, and monitoring data pipelines. It allows users to connect to various data sources, extract, transform, and load (ETL) data into their desired storage destinations. ADF is a serverless service, meaning that it is automatically provisioned and managed by Microsoft, significantly reducing the operational overhead for users.

Key Features of ADF

- **Codeless Development:** ADF offers a user-friendly graphical interface (UI) that allows users to create and manage data pipelines without writing code.
- **Integration with Various Data Sources:** ADF can connect to a wide range of data sources, including relational databases, cloud storage services, and SaaS applications.
- **ETL and Data Transformation:** ADF supports various data transformation techniques, including data wrangling, filtering, and cleansing.
- **Scalability and High Availability:** ADF is a scalable service that can handle large volumes of data and is designed to be highly available.
- **Cost-Effective:** ADF is a pay-as-you-go service, making it a cost-effective option for organizations of all sizes.

Benefits of Using ADF

- **Simplified Data Integration:** ADF simplifies the process of integrating data from disparate sources, making it easier to consolidate and analyze data.
- **Reduced Operational Costs:** ADF's serverless architecture and codeless development features significantly reduce the operational costs associated with traditional data integration solutions.
- **Improved Data Quality:** ADF's data transformation capabilities help to ensure the quality and consistency of data before it is loaded into target systems.
- **Centralized Management:** ADF provides a centralized platform for managing data pipelines, making it easier to monitor and troubleshoot data movement activities.

Overall, Azure Data Factory is a powerful and versatile data integration service that can help organizations of all sizes to streamline their data management processes.

Resource group - Project

In the resource group add service Azure data factory

The screenshot displays the Azure portal interface for the resource group 'bm-data-engineering-project'. The left sidebar shows the navigation menu with options like Overview, Activity log, Access control (IAM), Tags, Resource visualizer, Events, Settings, Deployments, Security, Deployment stacks, Policies, Properties, Locks, Cost Management, and Cost analysis. The main area shows the 'Essentials' section with a red box highlighting the 'Create' button. Below this, the 'Resources' section lists four resources: 'bm-data-engineering-KV' (Key vault), 'bm-Data-Engineering-project-ADF' (Data factory (V2)), 'bm-Data-Engineering-project-Databricks' (Azure Databricks Service), and 'bmdatalakegen2' (Storage account). The 'Search' bar at the top right shows the search term 'azure data factory' with a red box highlighting the search input. Below the search bar, the results show 1 to 20 of 61 results for 'azure data factory'. The results are displayed in a grid of cards, each representing a different service or solution. The first card is 'Excel Writer for Azure Data Factory' by Invari. The second card is 'Data Factory' by Microsoft, which is highlighted with a red box and a red arrow pointing to the 'Create' button. The third card is 'Delphix Compliance Services for Microsoft Azure' by Delphix. The fourth card is 'Cluedin Master Data Management' by Cluedin. The fifth card is 'Data#3 Azure Optimiser' by Data#3 Limited. The sixth card is 'Profisee SaaS Enterprise Master Data Management' by Profisee. The seventh card is 'Data#3 Azure Managed Services' by Data#3 Limited.

Name	Type	Location
bm-data-engineering-KV	Key vault	Poland Central
bm-Data-Engineering-project-ADF	Data factory (V2)	Poland Central
bm-Data-Engineering-project-Databricks	Azure Databricks Service	West Europe
bmdatalakegen2	Storage account	Poland Central

Showing 1 to 20 of 61 results for 'azure data factory'. [Clear search](#)

Showing 1 to 20 of 61 results for 'azure data factory'. [Clear search](#)

Showing 1 to 20 of 61 results for 'azure data factory'. [Clear search](#)

Create Data Factory ...

Basics Git configuration Networking Advanced Tags Review + create

One-click to create data factory with sample pipeline and datasets. [Try it](#)

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ	<input type="text" value="Azure subscription 1"/> ▼
└─ Resource group * ⓘ	<input type="text" value="bm-data-engineering-project"/> ▼
	Create new

Instance details

Name * ⓘ	<input type="text" value="data-engineer-project-adf-12"/> ✓
Region * ⓘ	<input type="text" value="East US"/> ▼
Version * ⓘ	<input type="text" value="V2"/> ▼

Previous

Next

Review + create

Azure Databricks

Azure Databricks is a fully managed cloud service for data engineering, data science, and machine learning. It combines the powerful Apache Spark engine with a streamlined user experience to help organizations of all sizes harness the power of their data.

Key Features of Azure Databricks

- **Unified Platform:** Azure Databricks provides a unified platform for data engineering, data science, and machine learning, making it easier to work with data across the entire data lifecycle.
- **Optimized Apache Spark Environment:** Azure Databricks provides an optimized Apache Spark environment that is pre-configured and managed by Microsoft. This makes it easy to spin up clusters and start working with data quickly.
- **Seamless Integration with Azure:** Azure Databricks is tightly integrated with other Azure services, such as Azure Storage, Azure Machine Learning, and Azure Synapse Analytics, making it easy to build and deploy data pipelines and machine learning models.
- **Scalable and High-Performance:** Azure Databricks is designed to be highly scalable and can handle large volumes of data. It is also designed for high-performance computing, making it ideal for complex data analysis tasks.
- **Cost-Effective:** Azure Databricks is a pay-as-you-go service, so you only pay for the resources you use. This makes it a cost-effective option for organizations of all sizes.

Benefits of Using Azure Databricks

- **Accelerated Data Engineering:** Azure Databricks can significantly accelerate the data engineering process by providing a streamlined user experience and pre-configured Apache Spark environment.
- **Enhanced Data Science Productivity:** Azure Databricks provides a rich set of tools and libraries for data science, making it easier to build and deploy machine learning models.
- **Simplified Machine Learning Deployment:** Azure Databricks integrates with Azure Machine Learning, making it easy to deploy machine learning models to production.
- **Improved Data Insights:** Azure Databricks can help organizations gain deeper insights from their data by enabling them to perform complex data analysis and machine learning tasks.

- **Reduced Data Costs:** Azure Databricks can help organizations reduce their data costs by making it easier to store, process, and analyze data in the cloud.

Overall, Azure Databricks is a powerful and versatile data analytics platform that can help organizations of all sizes harness the power of their data to gain insights, make better decisions, and drive business innovation.

Resource group - Project

In the resource group add service azure Databricks:

The screenshot shows the Azure portal interface. At the top, the resource group 'bm-data-engineering-project' is selected. A red box labeled '1' highlights the 'Essentials' section. Below this, a table lists resources within the group:

Name	Type	Location
bm-data-engineering-KV	Key vault	Poland Central
bm-Data-Engineering-project-ADF	Data factory (V2)	Poland Central
bm-Data-Engineering-project-Databricks	Azure Databricks Service	West Europe
bmdatalakegen2	Storage account	Poland Central

Below the table, a search bar contains the text 'databricks', highlighted with a red box labeled '2'. Below the search bar, a button labeled 'View suggestions' is visible. The search results show 1 to 20 of 59 results for 'databricks'. The first result is 'Azure Databricks' by Microsoft, which is highlighted with a red box labeled '3' over the 'Create' button.

Create an Azure Databricks workspace ...

Basics Networking Encryption Tags Review + create

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ

Azure subscription 1

Resource group * ⓘ

bm-data-engineering-project

Create new

Instance Details

Workspace name *

bricks

Region *

East US

Pricing Tier * ⓘ

Standard (Apache Spark, Secure with Microsoft Entra ID)

Managed Resource Group name

Enter name for managed resource group

Azure Key Vault

Azure Key Vault is a cloud-based service for securely storing and managing secrets, including passwords, certificates, and cryptographic keys. It is a highly secure service that uses hardware security modules (HSMs) to protect your secrets.

Why use Azure Key Vault?

There are many reasons to use Azure Key Vault, including:

- Improved security: Azure Key Vault uses HSMs to protect your secrets, which are more secure than storing them in your application code or database.
- Centralized management: Azure Key Vault provides a centralized location to store and manage all of your secrets, which makes it easier to track and audit them.
- Delegated access: You can control who has access to your secrets by using Azure Active Directory (Azure AD). This can help to prevent unauthorized access to your secrets.
- Reduced risk of breaches: By using Azure Key Vault, you can reduce the risk of data breaches caused by stolen or compromised secrets.

Key Vault features

Azure Key Vault offers a number of features that make it a powerful and versatile tool for managing secrets. These features include:

- Secret storage: Azure Key Vault can store a variety of secrets, including passwords, certificates, and cryptographic keys.
- Secret rotation: Azure Key Vault can automatically rotate your secrets on a regular schedule to help protect against key compromise.
- Access control: Azure Key Vault uses Azure AD to control who has access to your secrets. You can define granular access control policies to control who can read, write, and delete your secrets.
- Auditing: Azure Key Vault provides comprehensive auditing logs that you can use to track who accessed your secrets and when.
- Integration with other Azure services: Azure Key Vault can be integrated with other Azure services, such as Azure App Service and Azure Functions. This makes it easy to use your secrets in your applications.

How to use Azure Key Vault

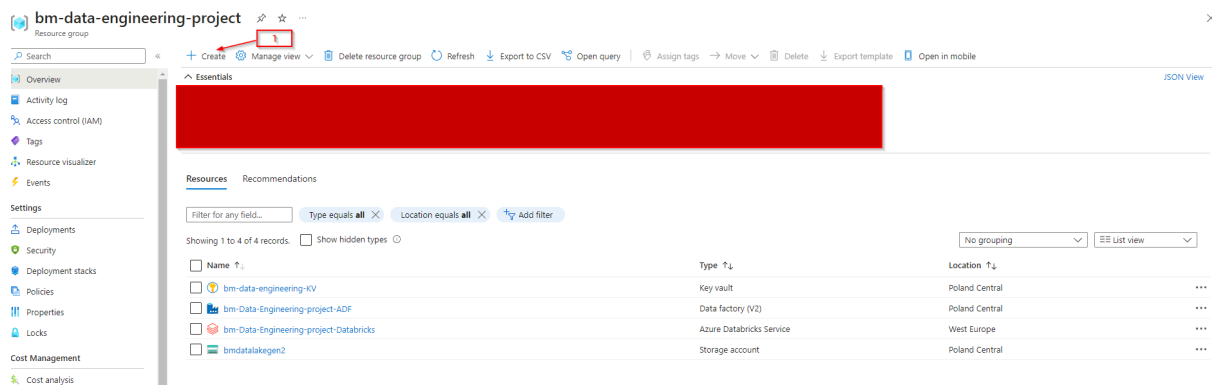
You can use Azure Key Vault to store and manage secrets in a number of ways, including:

- The Azure portal: You can use the Azure portal to create, manage, and access your secrets.
- The Azure CLI: You can use the Azure CLI to automate tasks related to Azure Key Vault.
- The .NET SDK: You can use the .NET SDK to integrate Azure Key Vault with your .NET applications.
- The REST API: You can use the REST API to interact with Azure Key Vault programmatically.

Overall, Azure Key Vault is a powerful and versatile service for securely storing and managing secrets. It is an essential part of any cloud security strategy.

Resource group - Project

In the resource group create Azure Key Vault service:



The screenshot shows the Azure portal interface for the resource group 'bm-data-engineering-project'. The 'Create' button is highlighted with a red box. Below the navigation pane, a table lists the resources in the group:

Name	Type	Location
bm-data-engineering-KV	Key vault	Poland Central
bm-Data-Engineering-project-ADF	Data factory (V2)	Poland Central
bm-Data-Engineering-project-Databricks	Azure Databricks Service	West Europe
bmdatalakegen2	Storage account	Poland Central

key vaults 2

Pricing : All

Operating System : All

Publisher Type : All

Product Type : All

Publisher name : All






☐ Azure services only



New! Get AI-generated suggestions for your search.

[View suggestions](#)

Showing 1 to 5 of 5 results for 'key vaults'. [Clear search](#)

<div></div> <div>Key Vault</div> <div>Microsoft</div> <div>Azure Service</div> <div>Safeguard cryptographic keys and other secrets used by cloud apps and services.</div> <div><div>3</div><div>Create</div></div>	<div></div> <div>Warren Averett Lighthouse Services</div> <div>Warren Averett Technology Group</div> <div>Managed Services</div> <div>Lighthouse managing service provides proactive & reactive support and maintenance for Azure.</div> <div><div></div><div>Create</div></div>	<div></div> <div>MNP LLP IT Managed Services</div> <div>MNP</div> <div>Managed Services</div> <div>Azure Lighthouse service for managed services customers of MNP Managed Services.</div> <div><div></div><div>Create</div></div>	<div></div> <div>Guard+ Standard</div> <div>KAMIND IT, Inc</div> <div>SaaS</div> <div>Kamind GUARD+ empowers IT Managers to plan Security Improvements</div> <div><div>Free trial</div><div>Subscribe</div></div>	<div></div> <div>Birlasoft Microservices Framework</div> <div>BIRLASOFT LIMITED</div> <div>Azure Application</div> <div>Structured & Automated Design and Implementation of Microservices-based Applications.</div> <div><div>Price varies</div><div>Create</div></div>
--	--	---	---	---

Create a key vault ...

Basics Access configuration Networking Tags Review + create

Azure Key Vault is a cloud service used to manage keys, secrets, and certificates. Key Vault eliminates the need for developers to store security information in their code. It allows you to centralize the storage of your application secrets which greatly reduces the chances that secrets may be leaked. Key Vault also allows you to securely store secrets and keys backed by Hardware Security Modules or HSMs. The HSMs used are Federal Information Processing Standards (FIPS) 140-2 Level 2 validated. In addition, key vault provides logs of all access and usage attempts of your secrets so you have a complete audit trail for compliance.

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *	<div>Azure subscription 1</div>
Resource group *	<div>bm-data-engineering-project</div> <div>Create new</div>

Instance details

Key vault name * ⓘ	<div>data-engineer-project-KV</div>
Region *	<div>East US</div>
Pricing tier * ⓘ	<div>Standard</div>

Recovery options

Soft delete protection will automatically be enabled on this key vault. This feature allows you to recover or permanently delete a key vault and secrets for the duration of the retention period. This protection applies to the key vault and the secrets stored within the key vault.

To enforce a mandatory retention period and prevent the permanent deletion of key vaults or secrets prior to the retention

<div>Previous</div>	<div>Next</div>	<div>Review + create</div>
---------------------	-----------------	----------------------------

Storage account

An Azure Storage Account is a cloud-based repository that stores data objects, including blobs, files, queues, and tables. It provides a unique namespace for your Azure Storage data that is accessible from anywhere in the world over HTTP or HTTPS.

Key features of Azure Storage Account:

- **Durability:** Azure Storage Account leverages redundancy and geo-replication to guarantee high availability and data durability.
- **Scalability:** Azure Storage Account can dynamically scale up or down to meet your changing storage needs.
- **Security:** Azure Storage Account employs various security measures to protect your data, including encryption, access control, and auditing.
- **Performance:** Azure Storage Account offers high performance for both read and write operations, making it suitable for a wide range of applications.
- **Cost-effectiveness:** Azure Storage Account offers a pay-as-you-go pricing model, allowing you to only pay for the resources you use.

Types of Azure Storage Account:

Azure Storage Account provides four primary types of data storage:

- **Blob storage:** Designed for storing unstructured data, such as images, videos, and documents.
- **File storage:** Provides a managed file share solution for cloud-based applications and on-premises file access.
- **Queue storage:** Optimized for storing and processing large numbers of messages in a reliable and ordered manner.
- **Table storage:** Efficiently stores structured data in a NoSQL format, ideal for applications that require fast access to large datasets.

Benefits of using Azure Storage Account:

- **Reduced infrastructure costs:** Eliminates the need for on-premises storage infrastructure, saving you hardware, software, and maintenance costs.
- **Scalability and flexibility:** Easily scale your storage capacity up or down to meet your changing needs, without upfront investments.
- **High availability and durability:** Ensures your data is protected from failures and disasters, ensuring business continuity.

- Security and compliance: Employs robust security measures to protect your data, meeting compliance requirements for various industries.
- Global reach: Access your data from anywhere in the world with low latency thanks to Azure's global network of data centers.

Overall, Azure Storage Account is a versatile and scalable cloud storage solution that can meet the needs of a wide range of organizations. Its durability, scalability, security, and cost-effectiveness make it an ideal choice for storing and managing essential data.

Resource group - Project

The screenshot shows the Azure Marketplace search results for 'storage account'. At the top, there is a search bar with 'storage account' entered, and a red box labeled '1' highlights the search bar. Below the search bar, there are filters for Pricing, Operating System, Publisher Type, Product Type, and Publisher name, all set to 'All'. A checkbox for 'Azure services only' is also present. A banner below the filters says 'New! Get AI-generated suggestions for your search.' with a 'View suggestions' button. Below the banner, it says 'Showing 1 to 20 of 180 results for 'storage account''. A 'Tile view' dropdown is on the right. The results are displayed as a grid of seven tiles. The first tile is 'Storage account' by Microsoft, with a red box labeled '2' and an arrow pointing to the 'Create' button. The other tiles include 'Storage Account Using ARM Template' by FortuneCloud LLC, 'Azure Storage Mover' by Microsoft, 'Storage Account Using ARM' by DIGISTORM LTD., 'Storage Account Using ARM' by VIRTUCLLOUD LTD., 'MDACA Cloud Storage Explorer' by Spin Systems Inc, and 'APEX Protection Storage for Microsoft Azure (DDVE)' by Dell Technologies.

Create a storage account ...

[Basics](#) [Advanced](#) [Networking](#) [Data protection](#) [Encryption](#) [Tags](#) [Review](#)

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription *	<div>Azure subscription 1</div>
Resource group *	<div>bm-data-engineering-project</div> <div>Create new</div>

Instance details

Storage account name ⓘ *	<div>bmdatalakegen21</div>
Region ⓘ *	<div>(Europe) Poland Central</div> <div>Deploy to an edge zone</div>
Performance ⓘ *	<div><input checked="" type="radio"/> Standard: Recommended for most scenarios (general-purpose v2 account)</div> <div><input type="radio"/> Premium: Recommended for scenarios that require low latency.</div>
Redundancy ⓘ *	<div>Locally-redundant storage (LRS)</div>

3

[Review](#)

[< Previous](#)

[Next : Advanced >](#)

In th resource group create following service:

1. ADF
2. Azure Databricks
3. Key vault
4. Storage account
5. Synapse workspace

Set up Azure Active Directory

Data source - on-premise database

Data source: SMSS - SQL server


Data source will be taken from the Microsoft sample database the AdventureWorks - AdventureWorksLT2022.bak

Download backup files

Use these links to download the appropriate sample database for your scenario.

- OLTP data is for most typical online transaction processing workloads.
- Data Warehouse (DW) data is for data warehousing workloads.
- Lightweight (LT) data is a lightweight and pared down version of the OLTP sample.

If you're not sure what you need, start with the OLTP version that matches your SQL Server version.

 Expand table

OLTP	Data Warehouse	Lightweight
AdventureWorks2022.bak	AdventureWorksDW2022.bak	AdventureWorksLT2022.bak
AdventureWorks2019.bak	AdventureWorksDW2019.bak	AdventureWorksLT2019.bak
AdventureWorks2017.bak	AdventureWorksDW2017.bak	AdventureWorksLT2017.bak
AdventureWorks2016.bak	AdventureWorksDW2016.bak	AdventureWorksLT2016.bak
AdventureWorks2016_EXT.bak	AdventureWorksDW2016_EXT.bak	N/A
AdventureWorks2014.bak	AdventureWorksDW2014.bak	AdventureWorksLT2014.bak
AdventureWorks2012.bak	AdventureWorksDW2012.bak	AdventureWorksLT2012.bak
AdventureWorks2008R2.bak	AdventureWorksDW2008R2.bak	N/A

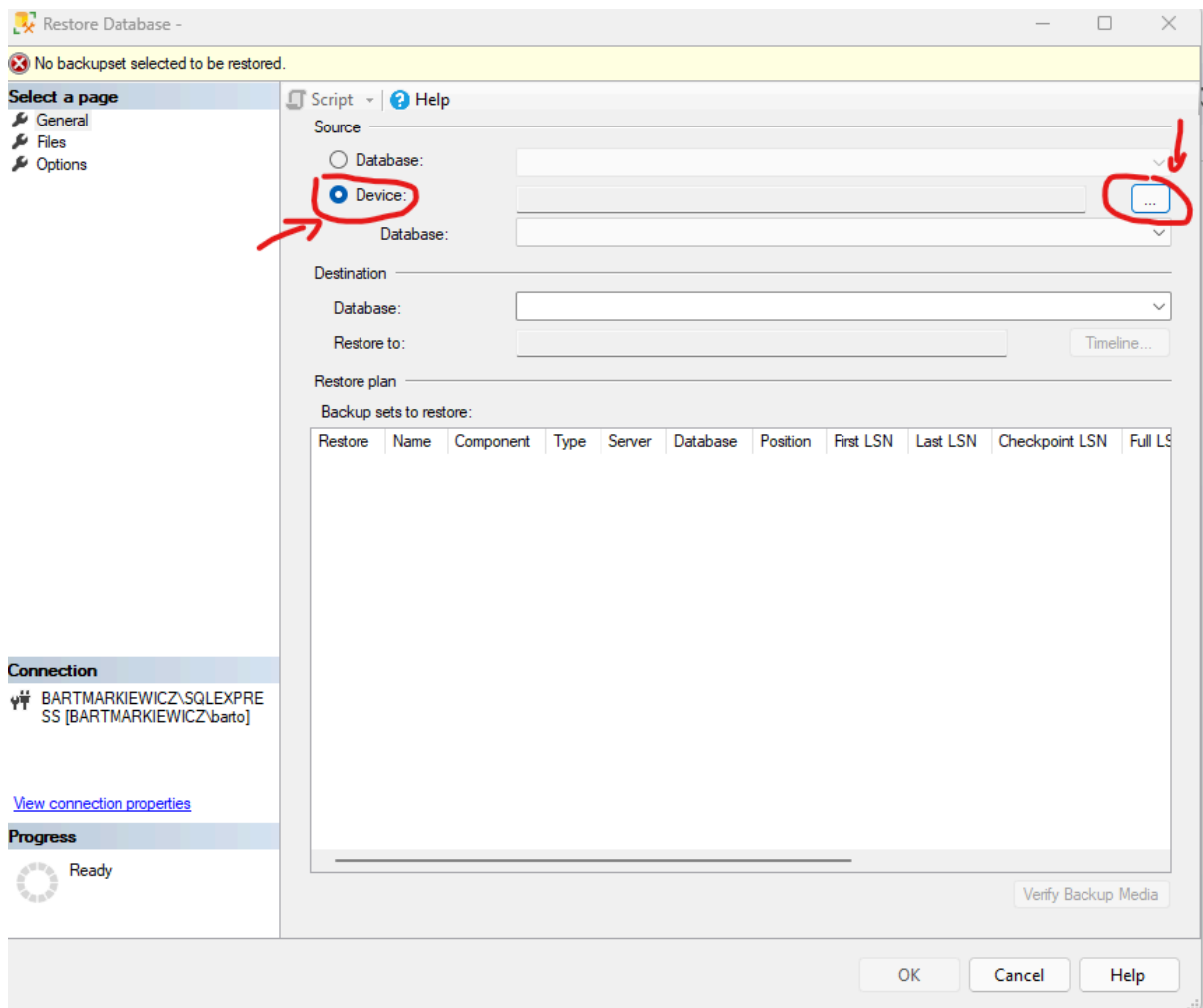
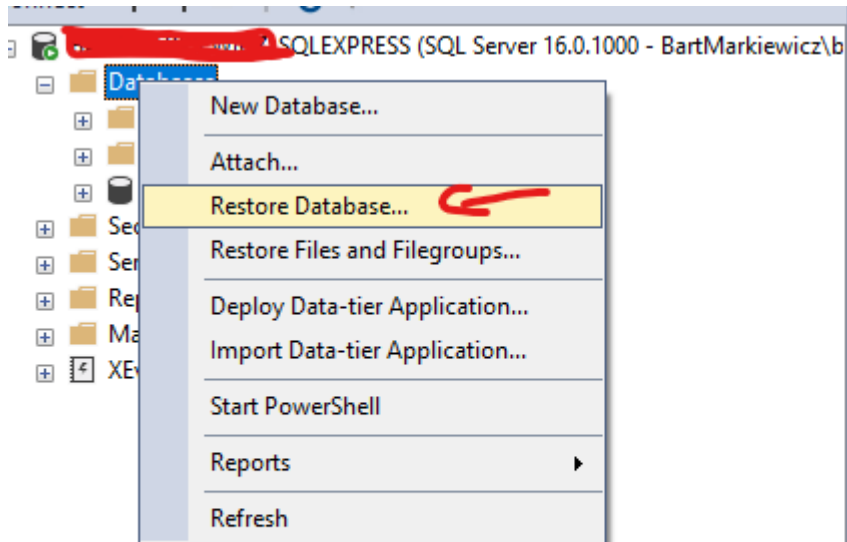
Additional files can be found directly on GitHub:

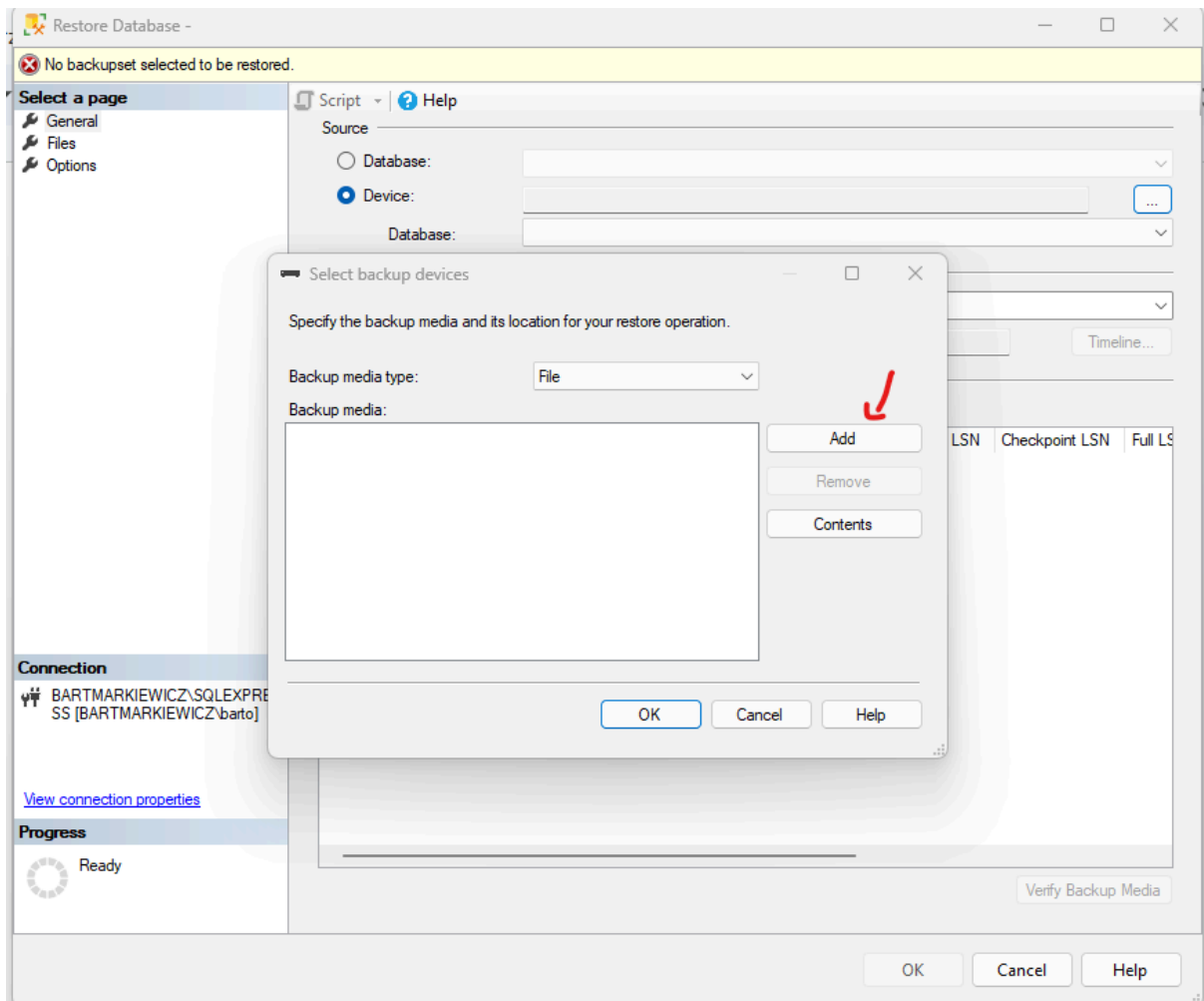
Link below:

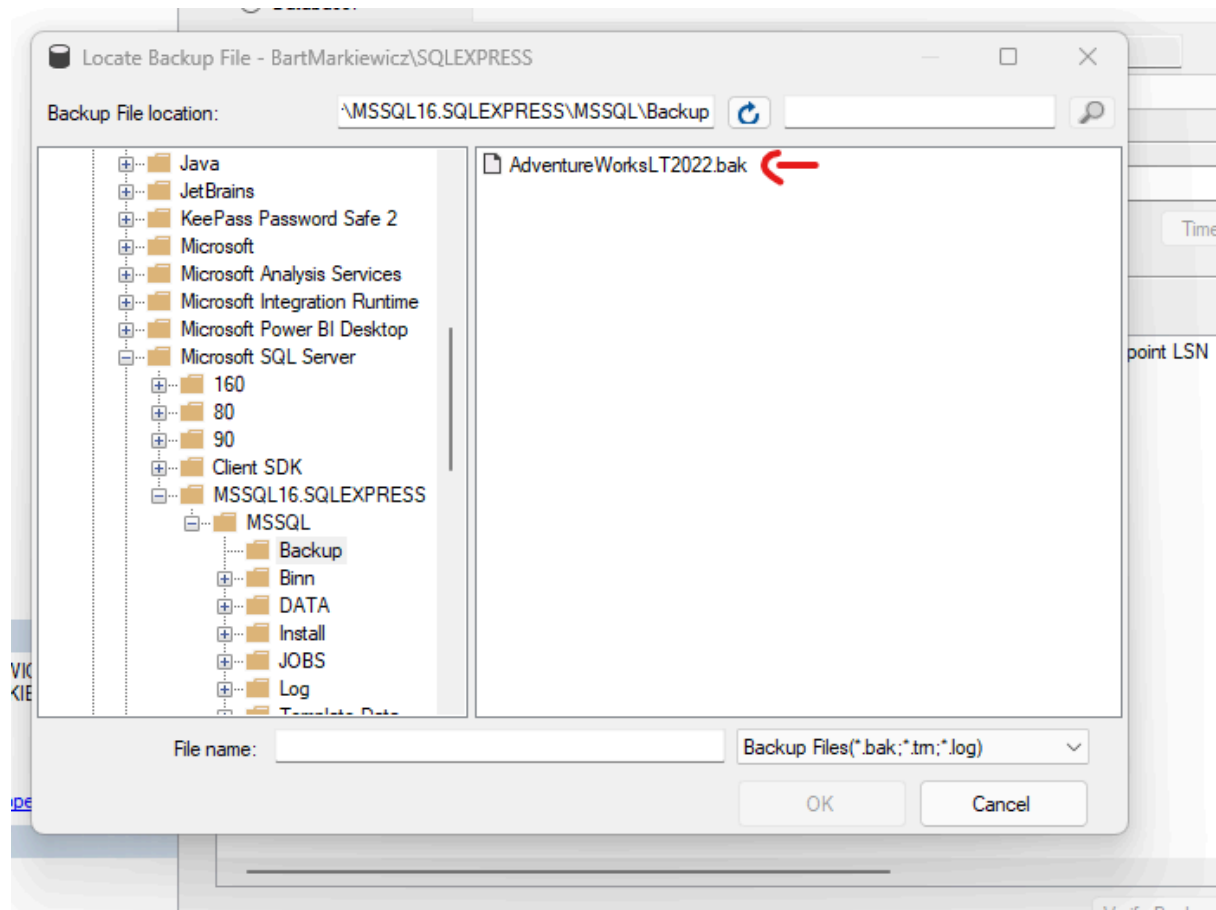
<https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms>

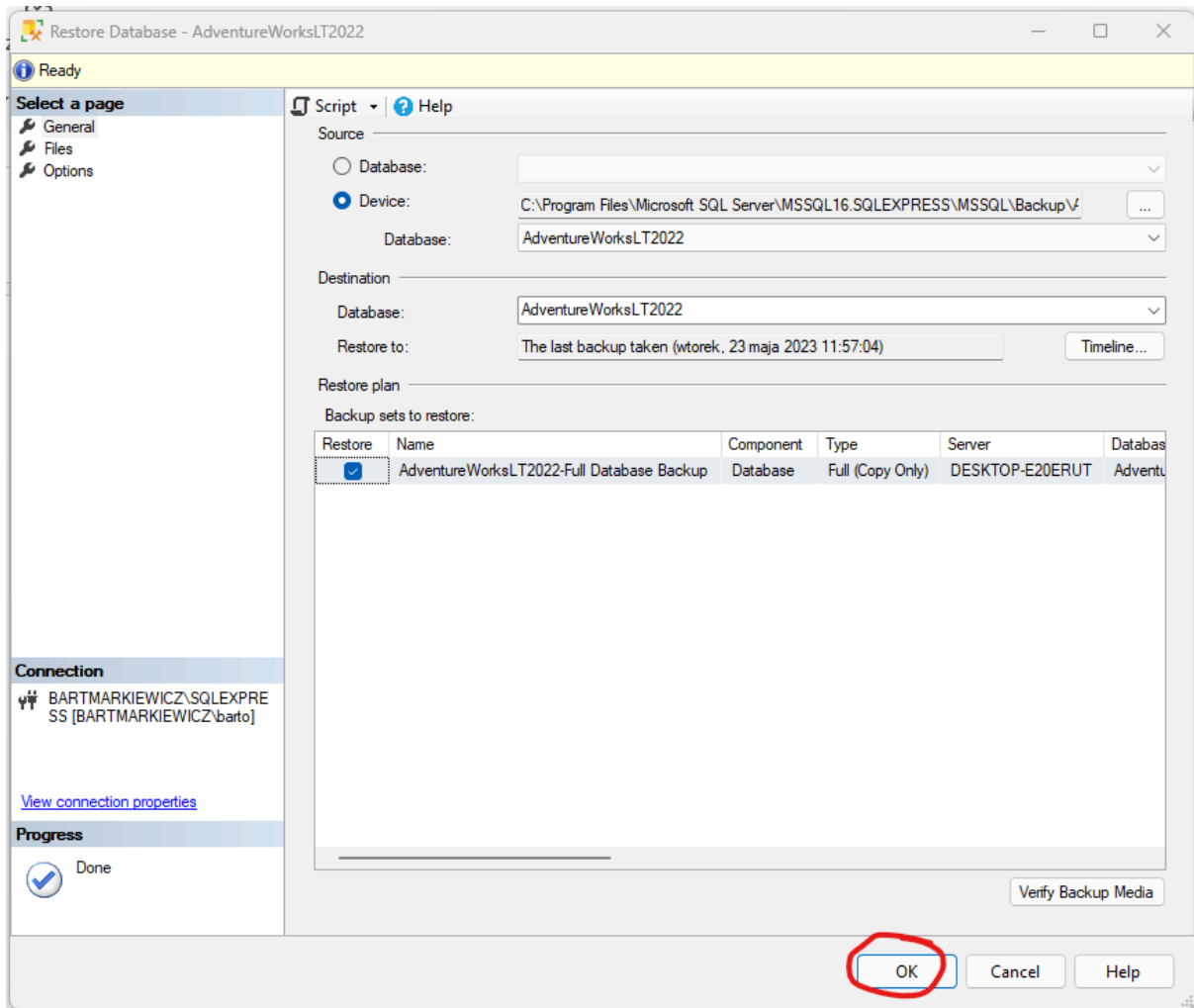
SQL - server configuration:

Load the backup of the downloaded database (before you started load the backup to the database, it might be necessary to move your download backup of database to the folder of smss to following path: MSSQL16.SQLEXPRESS\MSSQL\Backup (backup folder of MSSQL))









SQLSERVER (SQL Server 16.0.1000 - BartMarkiewicz\b

Databases

System Databases

Database Snapshots

AdventureWorksLT2022

Database Diagrams

Tables

Views

External Resources

Synonyms

Programmability

Service Broker

Storage

Security

Security

Server Objects

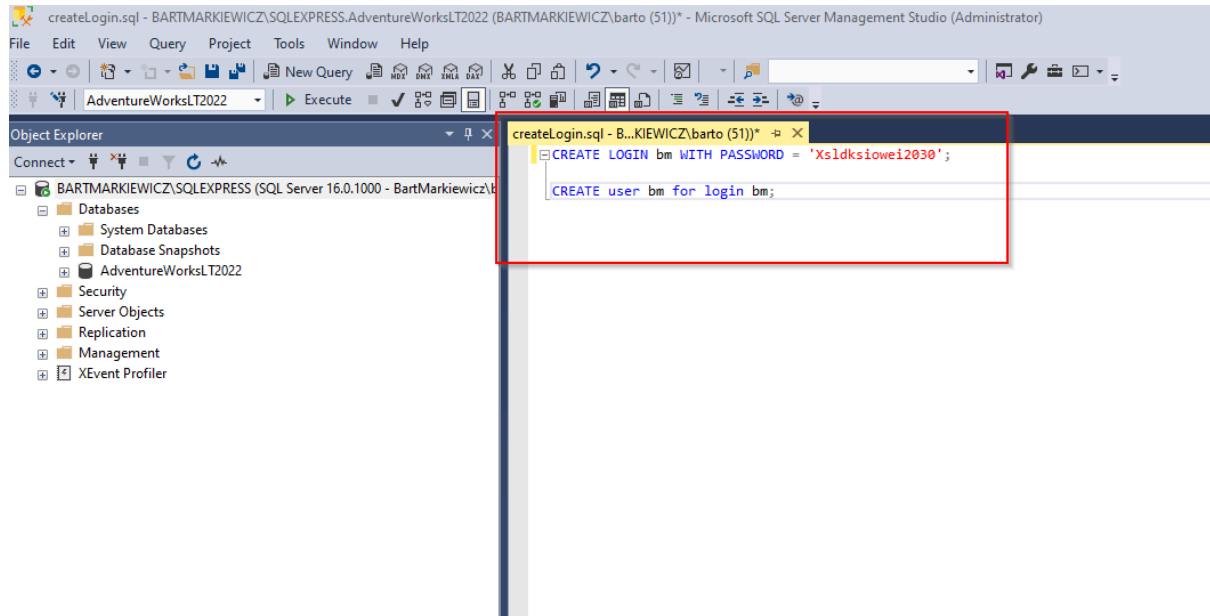
Replication

Management

XEvent Profiler

Create Account and user

Create an Account to the database AdventureWorksLT2022 then create user



Data Ingestion: Azure data factory - setup

Configuration run Integration Service

To connect Azure Data Factory with the on-premise you need to create Integration runtimes, that allow you to connect those two services.

To create "Integration runtimes" you need to go:

1. Manage
2. Integration runtime
3. Click new
4. Azure Self-Hosted
5. Self-Hosted
6. Type name of the service (in my case is: Project-DataEngineer)
7. Choose express setup - thankfully this option you don't have to manual install the service

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Home Author Monitor **Manage** Learning Center

Data Factory Factory settings Linked services **Integration runtimes** Microsoft Purview Source control Git configuration ARM template Author Triggers Global parameters Data flow libraries Security

Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. [Learn more](#)

+ New Refresh

Filter by name

Showing 1 - 1 of 1 items

Name	Type	Sub-type	Status	Related	Region
AutoResolveIntegrationR...	Azure	Public	Running	0	Auto Resolve

Data Factory bn-Data-Engineering-project-ADF Search factory and documentation bartosz.markiewicz01@gmail.com

Integration runtime setup

Integration Runtime is the native compute used to execute or dispatch activities. Choose what integration runtime to create based on required capabilities. [Learn more](#)

Azure, Self-Hosted
Perform data flows, data movement and dispatch activities to external compute.

Azure-SSIS
Lift-and-shift existing SSIS packages to execute in Azure.

Continue Cancel

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click [here](#) to get started with Fabric Data Factory!

Home Author Monitor **Manage** Learning Center

Data Factory Factory settings Linked services **Integration runtimes** Microsoft Purview Source control Git configuration ARM template Author Triggers Global parameters Data flow libraries Security Credentials Customer managed key Outbound rules

Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment.

+ New Refresh

Filter by name

Showing 1 - 1 of 1 items

Name	Type	Sub-type	Status
AutoResolveIntegrationR...	Azure	Public	Running

Integration runtime setup

Network environment:
Choose the network environment of the data source / destination or external compute to which the integration runtime will connect to for data flows, data movement or dispatch activities:

Azure
Use this for running data flows, data movement, external and pipeline activities in a fully managed, serverless compute in Azure.

Self-Hosted
Use this for running activities in an on-premises / private network. [View more](#)

External Resources:
You can use an existing self-hosted integration runtime that exists in another resource. This way you can reuse your existing infrastructure where self-hosted integration runtime is setup.

Linked Self-Hosted
[Learn more](#)

Microsoft Azure

Data Factory

1m-Data-Engineering-project-ADF

Search factory and documentation

bartosz.markiewicz01@gmail.com

DEFAULT DIRECTORY

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Home

Author

Monitor

Manage

Learning Center

General

Factory settings

Connections

Linked services

Integration runtimes

Microsoft Purview

Source control

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

Credentials

Customer managed key

Outbound rules

Managed private endpoints

Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environments.

+ New Refresh

Filter by name

Showing 1 - 1 of 1 items

Name	Type	Sub-type	Status
AutoResolveIntegrationR...	Azure	Public	Running

Integration runtime setup

Private network support is realized by installing integration runtime to machines in the same on-premises network/VNET as the resource the integration runtime is connecting to. Follow below steps to register and install integration runtime on your self-hosted machines.

NameProject-DataEngineering

DescriptionUsed to connect with SQL Server

TypeSelf-Hosted

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Home

Author

Monitor

Manage

Learning Center

General

Factory settings

Connections

Linked services

Integration runtimes

Microsoft Purview

Source control

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

Credentials

Customer managed key

Outbound rules

Managed private endpoints

Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environments.

+ New Refresh

Filter by name

Showing 1 - 1 of 1 items

Name	Type	Sub-type	Status
AutoResolveIntegrationR...	Azure	Public	Running

Integration runtime setup

SettingsNodesAuto updateSharingLinks

Install integration runtime on Windows machine or add further nodes using the Authentication Key.

NameProject-DataEngineering

Option 1: Express setup

Click here to launch the express setup for this computer

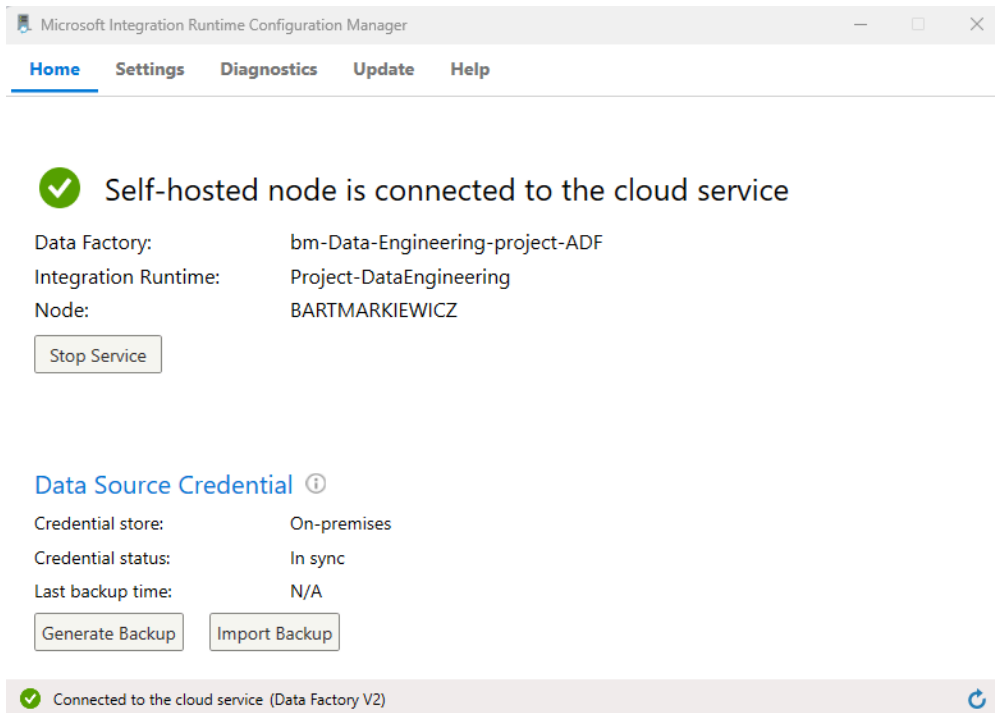
Option 2: Manual setup

Step 1: Download and install integration runtime

Step 2: Use this key to register your integration runtime

Name	Authentication key
Key1	
Key2	

If you correctly setup/configure your Integration runtime service you should see the following output in the Integration Runtime service

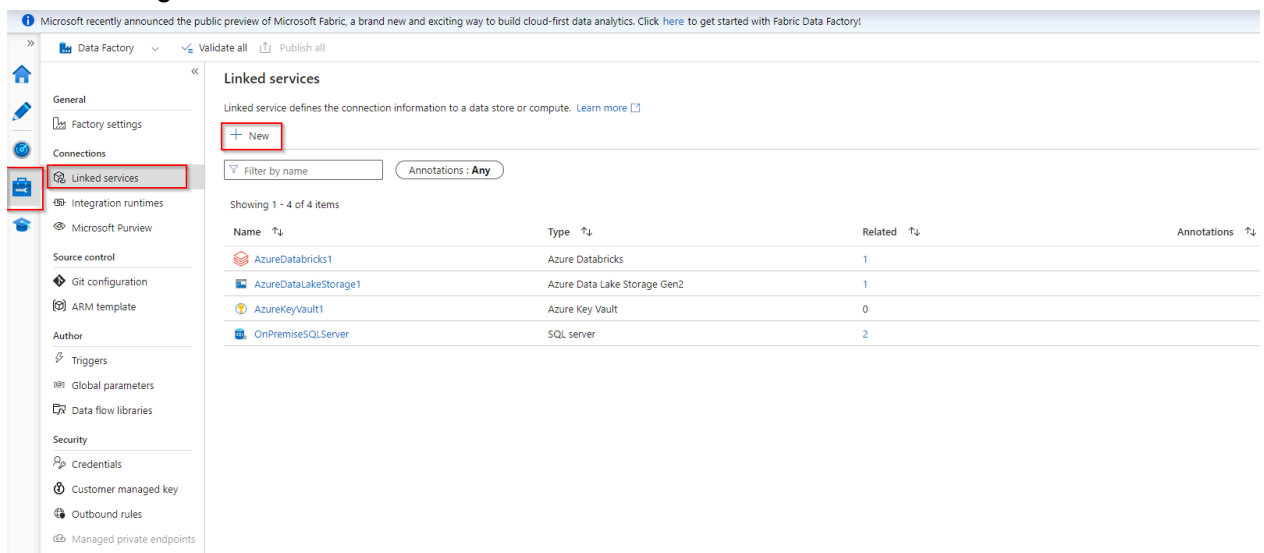


Configuration Linked services

After specified integration runtimes, you need to specify linked services, where you define the connection information to a data store or compute.

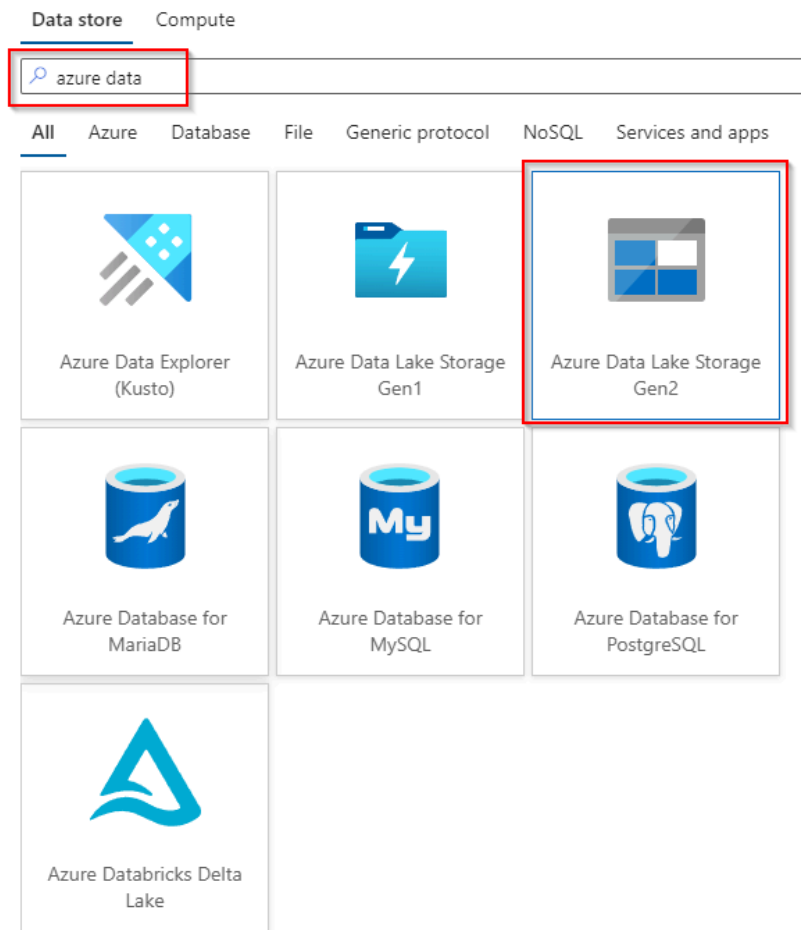
We will specify linked services to “Azure Data Lake Storage Gen2”

1. Go to manage, then Linked services and click “New”



2. Select “Azure Data Lake Storage Gen2”

New linked service



3. Specify:

- Name
- Connect via integration runtime setup as the `AutoResolveIntegrationRuntime`
- Authentication type, specify as: Account key
- Azure subscription: select your subscription
- Storage account name: select your storage account
- Test connection: To linked service
- Click test connection to check if you are able to connect to the service


New linked service

 Azure Data Lake Storage Gen2 [Learn more](#) 

Name *

AzureDataLakeStorage2

Description

Connect via integration runtime * 


AutoResolveIntegrationRuntime

Authentication type

Account key

Account selection method 

☒ From Azure subscription ☐ Enter manually

Azure subscription 

Azure subscription 1

Storage account name *

bmdatalakegen2

Test connection 

☒ To linked service ☐ To file path

Annotations


+ New

> Parameters

> Advanced 

Create

Back

 Test connection

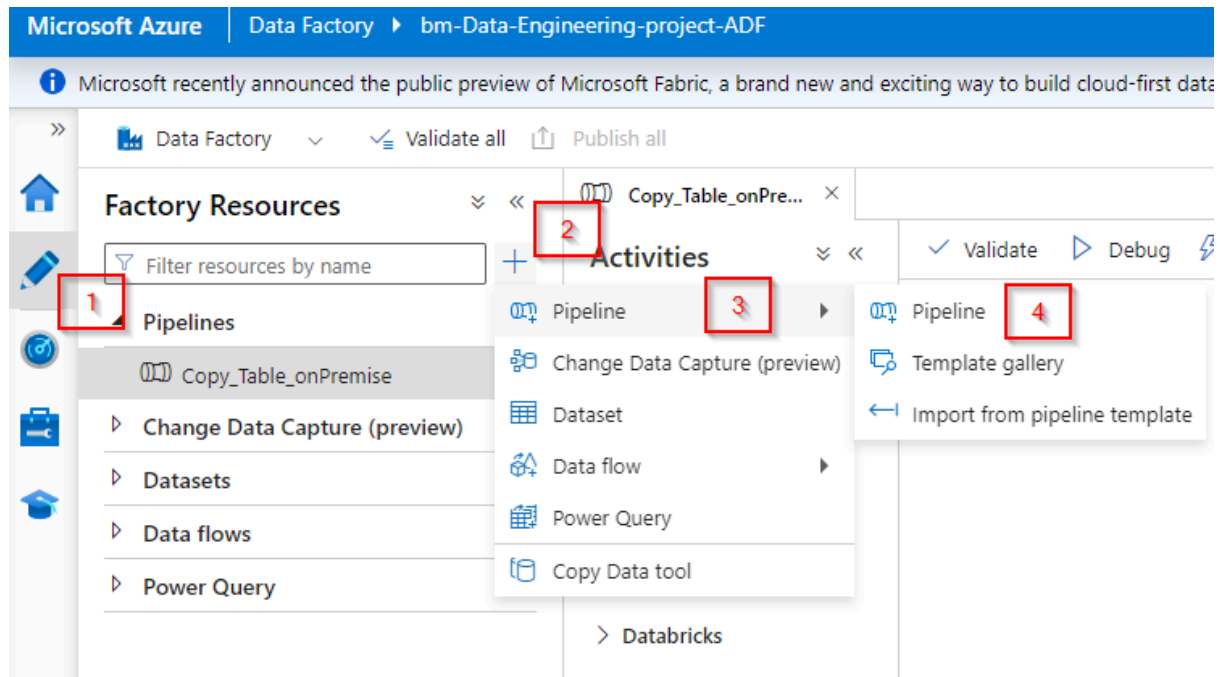
Cancel

Configuration Pipeline

The prepared pipeline will iterate through all tables in our database and take all tables with schema "SalesLT", after that all tables with schema "SalesLT" will be store in our azure storage in the folder bronze, then by using data bricks we will conduct transformations for layers silver and gold.

To create a pipeline you need to:

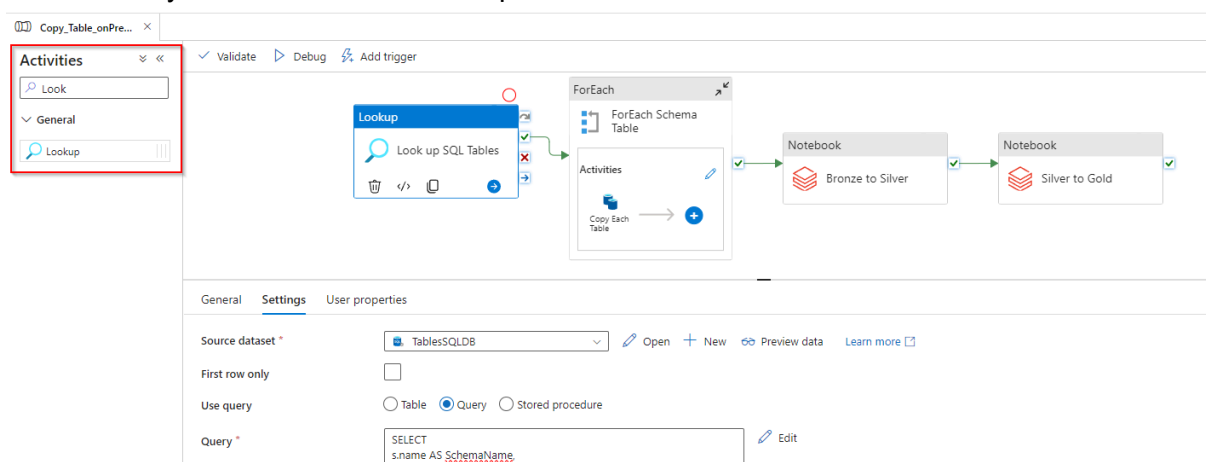
1. Go to manage
2. Click “+”
3. Pipeline → Pipeline
4. Name Pipeline (in the project is entitled as “Copy_table_onPremise”)



In the pipeline will be specified a couple of activities:

1. Lookup
2. ForEach
3. Notebook (allow us to execute pyspark code from Azure bricks)

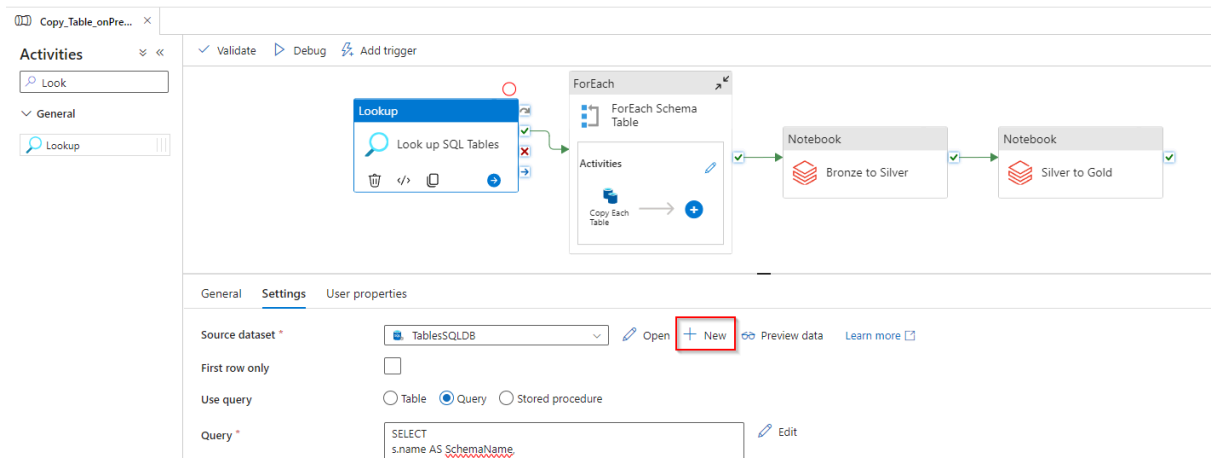
All Activities you should find in the left panel



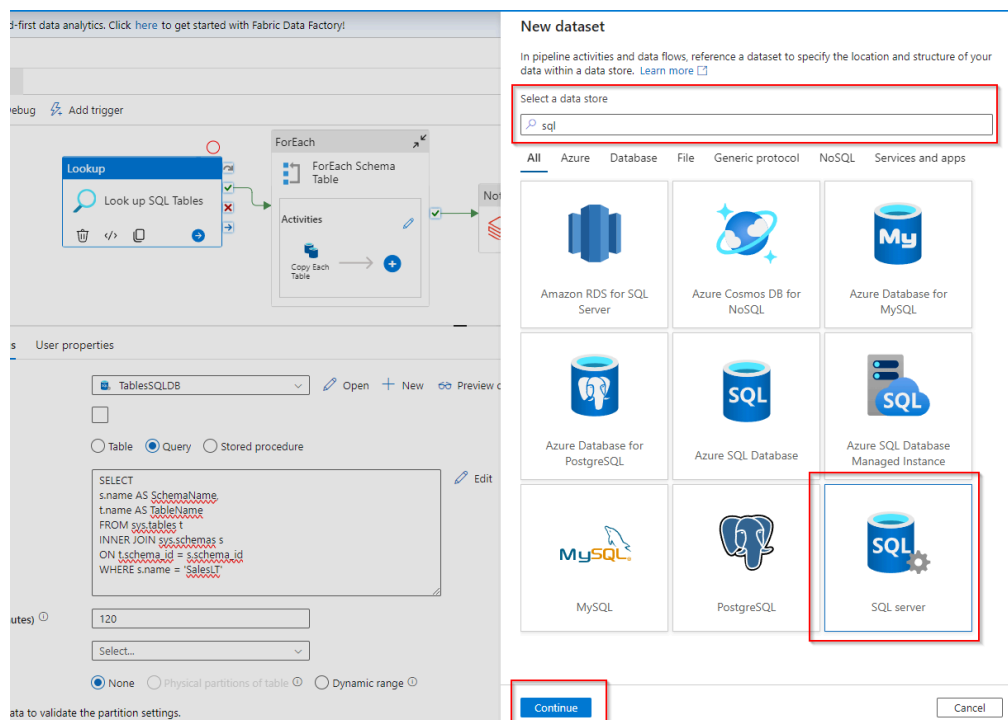
Lookup

In the Lookup activity, we will specify the tables, which we would like to copy to our storage account. To do that you need:

1. Click on Lookup activity
2. In General type the name of the activities (in the project I used “Look up SQL Tables”)
3. Go to Settings
4. Create new dataset (Click new “+”)



5. Select SQL server



6. Type the name of the dataset (in my case I used TablesSQLDB) and important do not specify table name

Set properties

Name
TablesSQLDB1

Linked service *
OnPremiseSQLServer

Connect via integration runtime * ⓘ
Project-DataEngineering

Table name
Select...
☐ Enter manually

Import schema
☐ From connection/store ☒ None

> Advanced

OK Back Cancel

7. After creating dataset you should be able to select source in the lookup activities
8. In the Query panel you will use the query to list all tables with schema "SalesLT"

General Settings User properties

Source dataset *
First row only
Use query
Query *

TablesSQLDB
OnPremise_Address
parquet OnPremise_Address
TablesSQLDB

SELECT
s.name AS SchemaName
t.name AS TableName
FROM sys.tables t
INNER JOIN sys.schemas s
ON t.schema_id = s.schema_id
WHERE s.name = 'SalesLT'

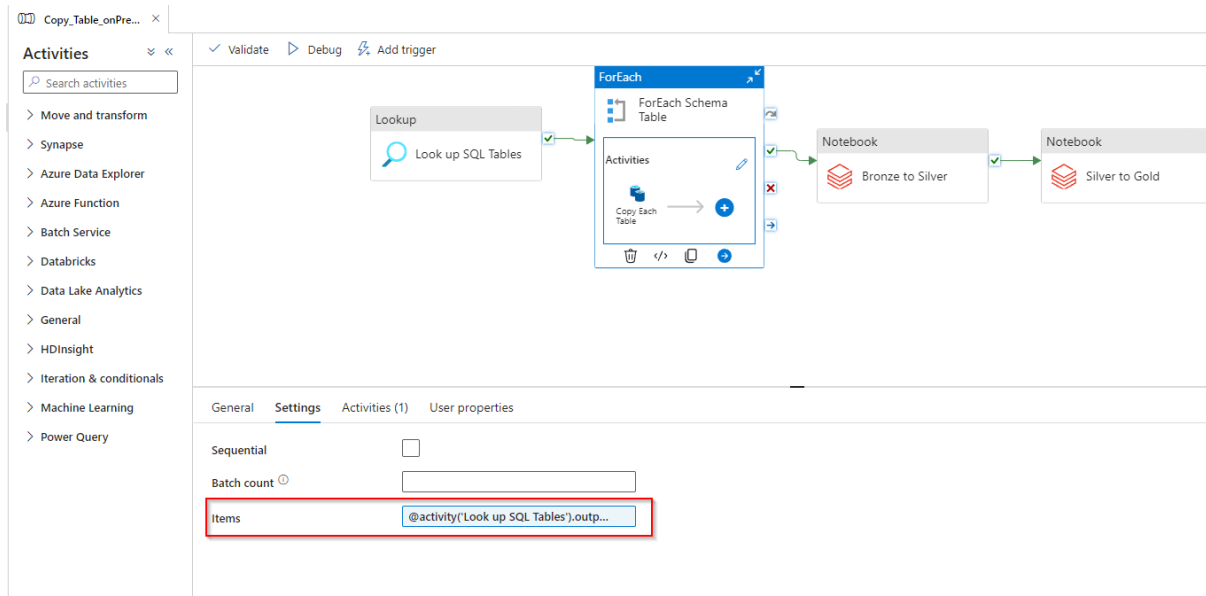
Query timeout (minutes) ⓘ 120
Isolation level ⓘ Select...

SQL Query:

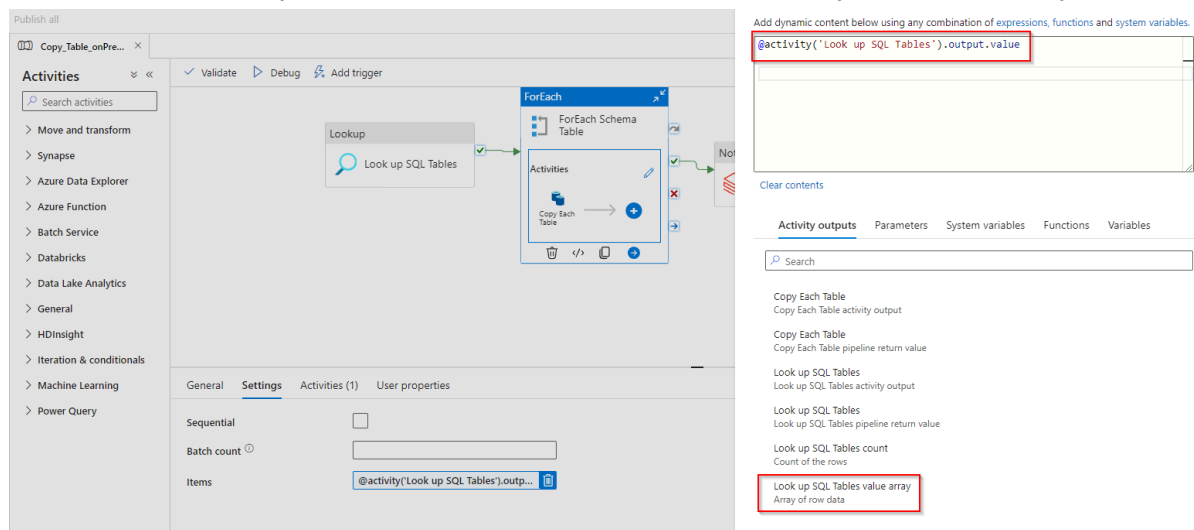
```
SELECT
s.name AS SchemaName,
t.name AS TableName
FROM sys.tables t
INNER JOIN sys.schemas s
ON t.schema_id = s.schema_id
WHERE s.name = 'SalesLT'
```

ForEach

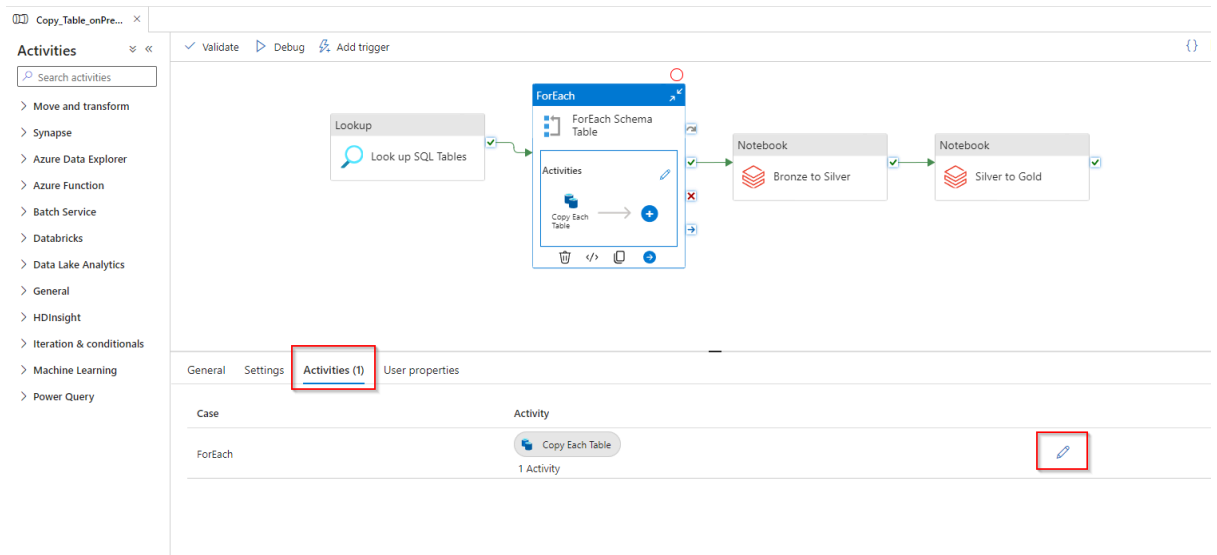
1. Create new activities forEach
2. Go to settings → click items



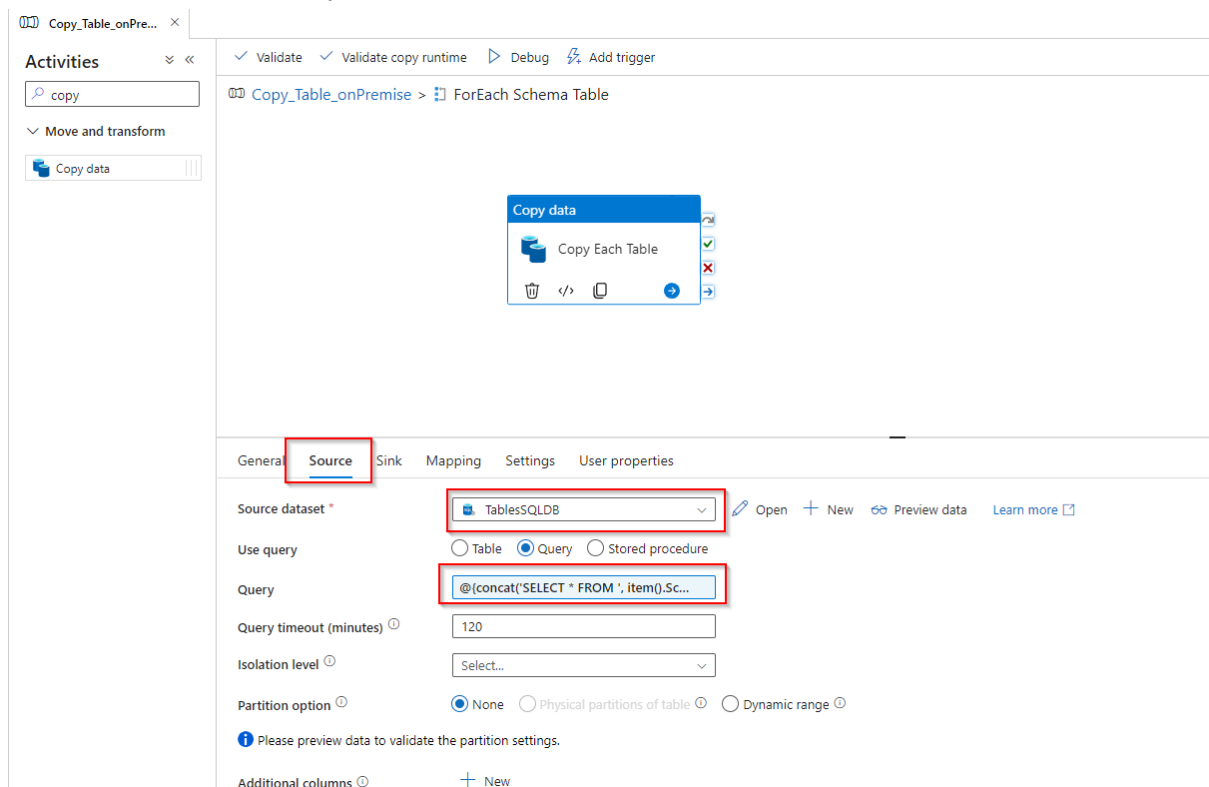
3. Select “Activity outputs” - Look up SQL Tables value array / You can also type



4. In the “Activities” section edit forEach activity

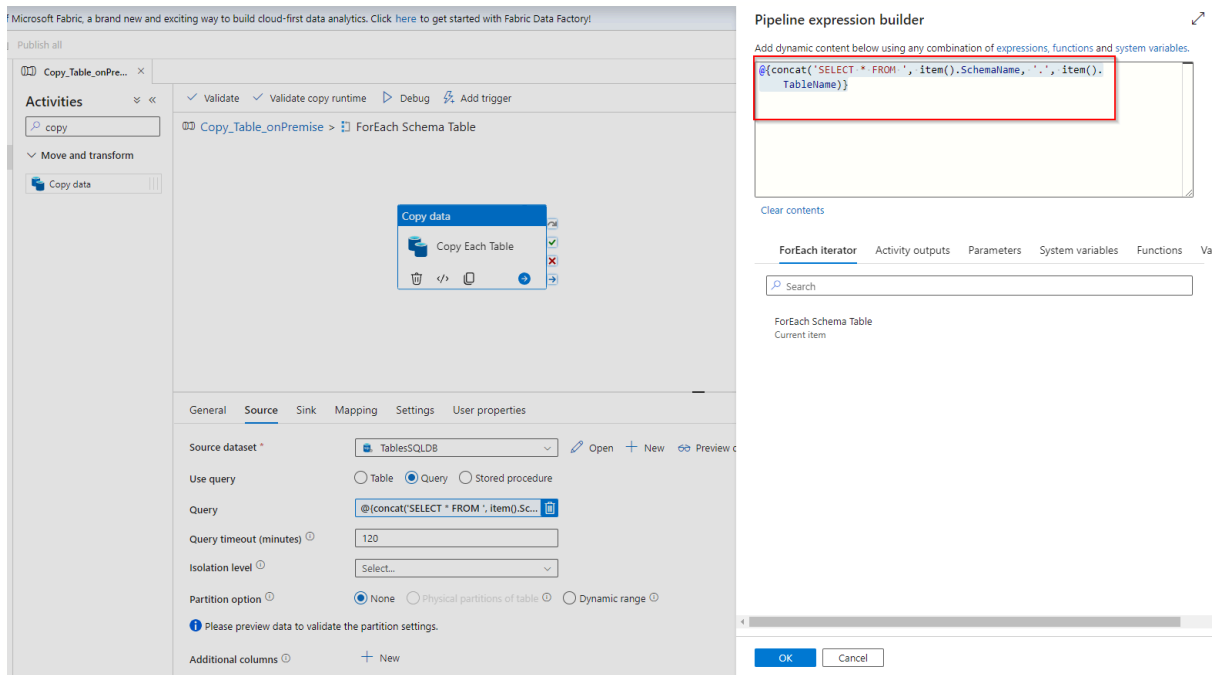


5. Take the activity “Copy data”
6. In the “Copy data” activity go to Source
7. In the source select the source (the ones that has been prepared before)
8. Click on the Query

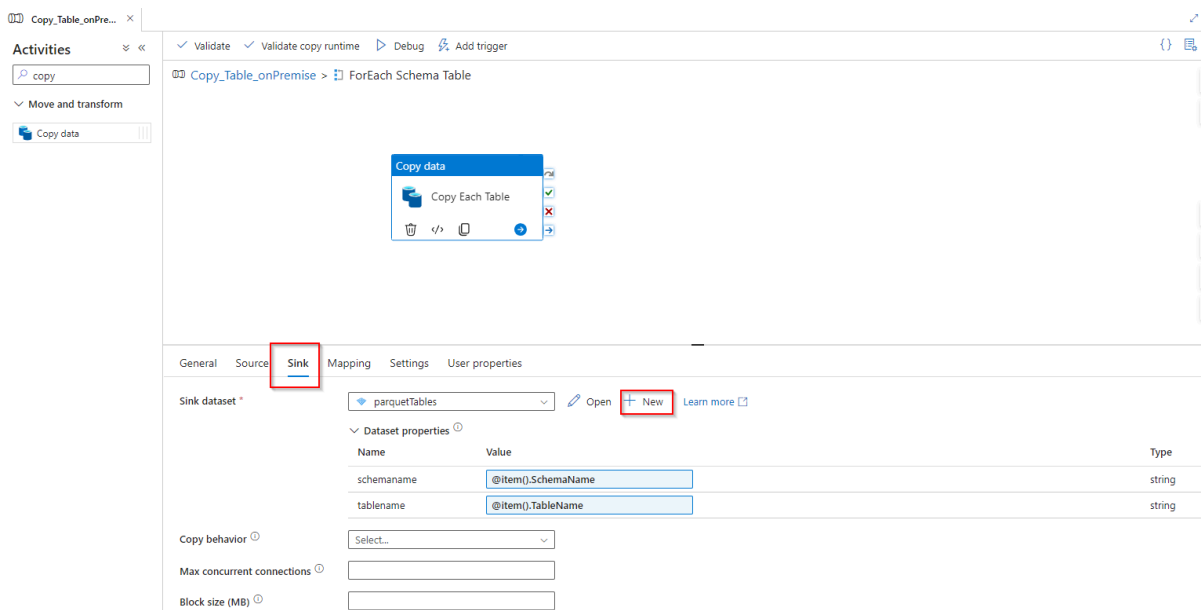


9. Type in the “Pipeline expression builder” following expression:

```
@{concat('SELECT * FROM ', item().SchemaName, '...', item().TableName)}
```

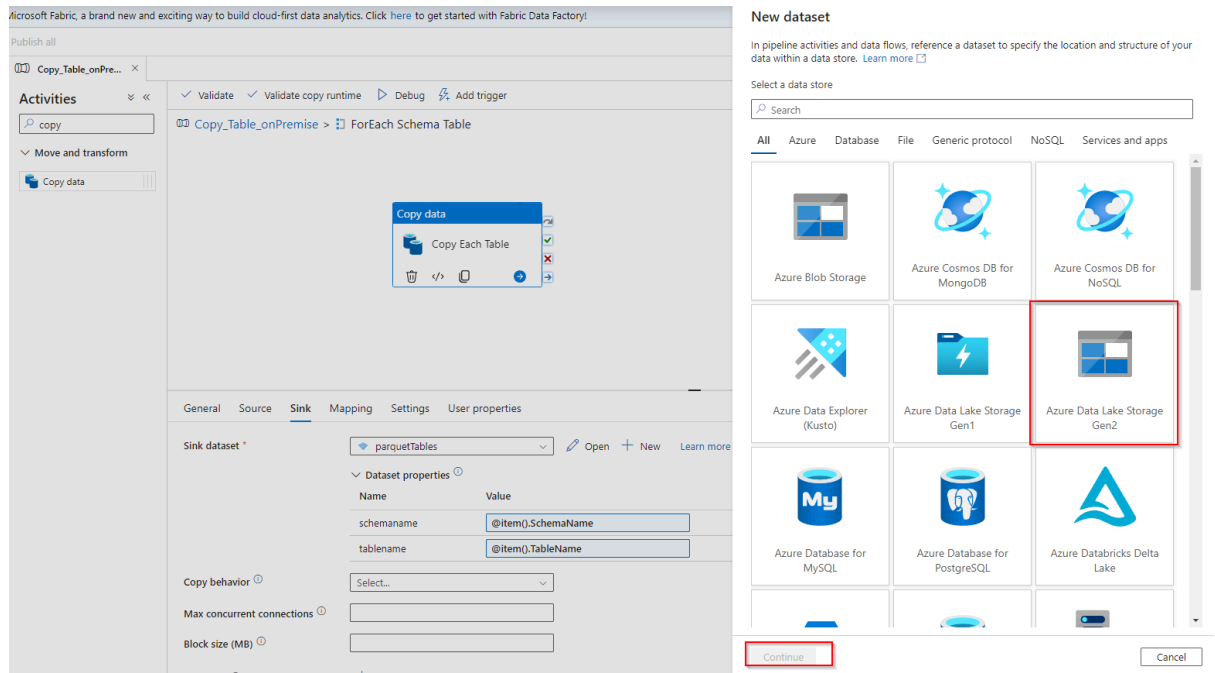


10. Go to Sink section

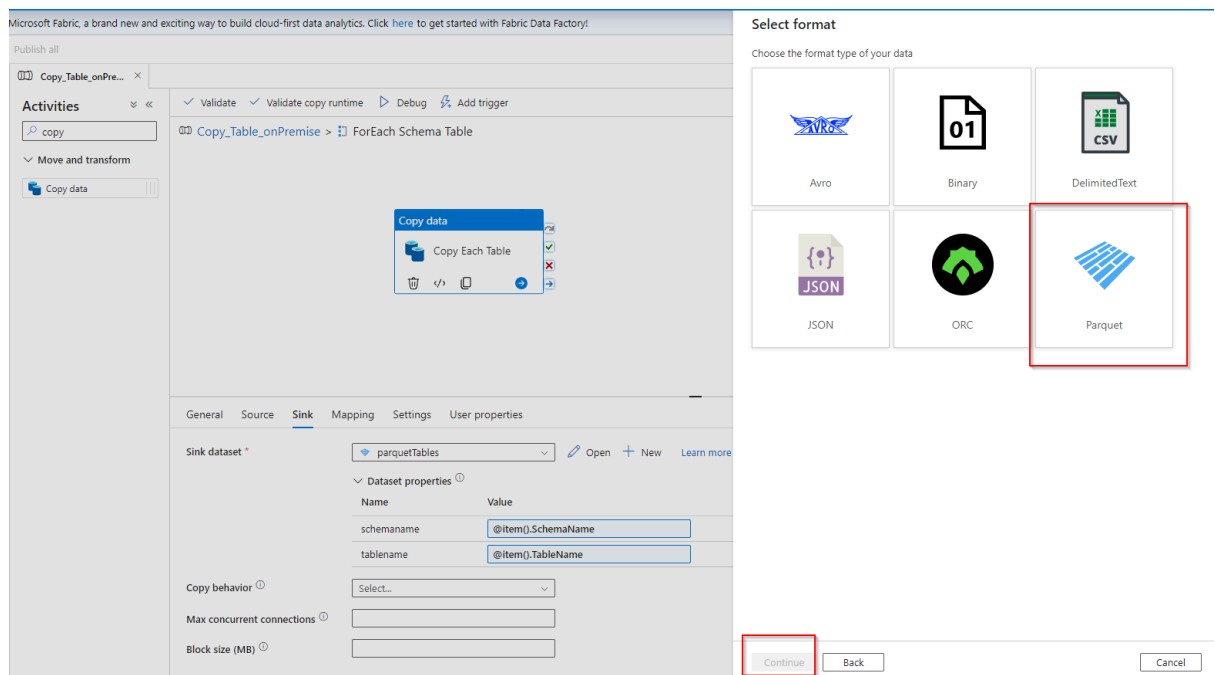


11. Create new dataset

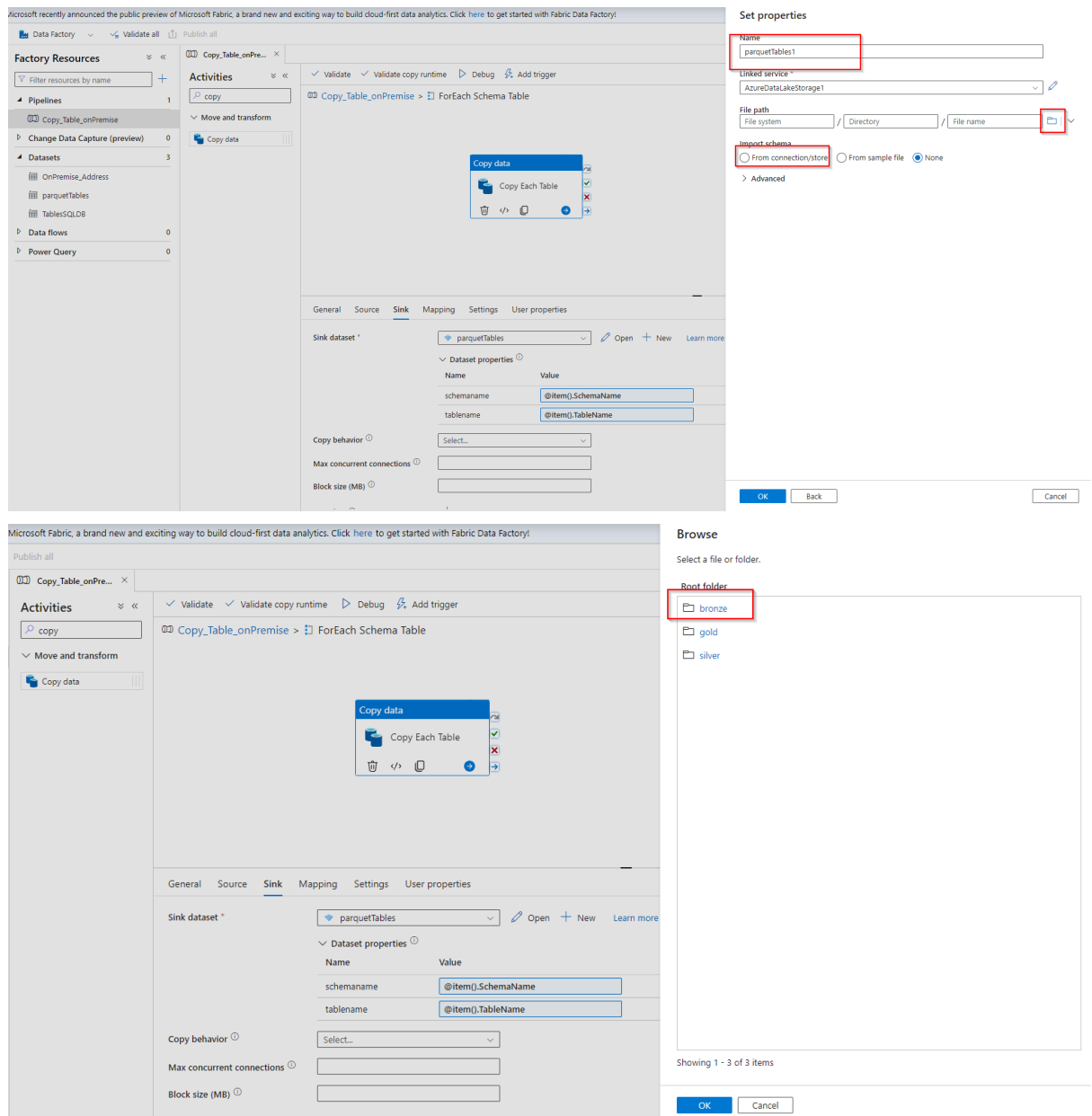
12. Select Azure Data lake Storage Gen2 → continue



13. Select parquet



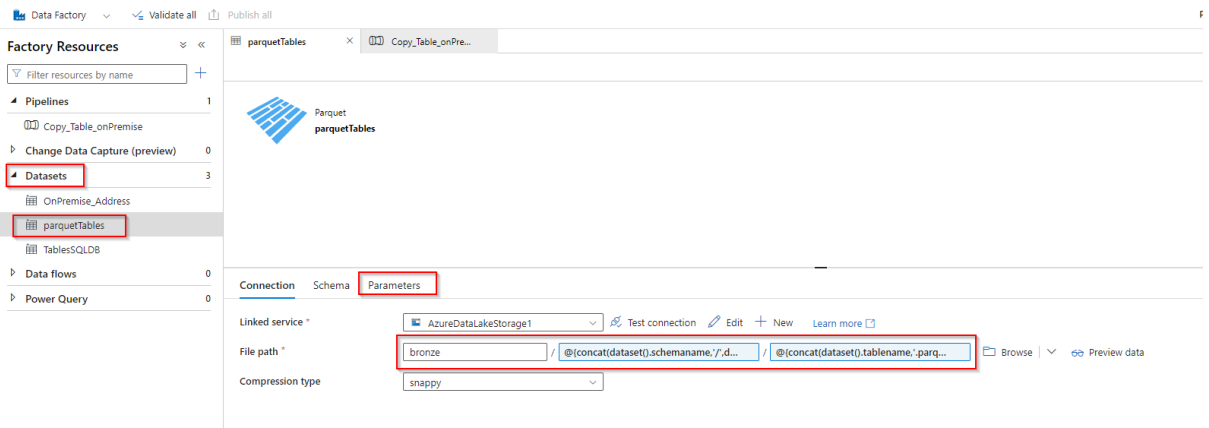
14. Select bronze folder (before, you need to create in Storage account service, three containers - bronze / silver / gold), directory and file name leave blank. The structure of the file, which will be used in the project is bronze/Schema/Tablename/Tablename.parquet example:
bronze/SalesLT/Address/Address.parquet



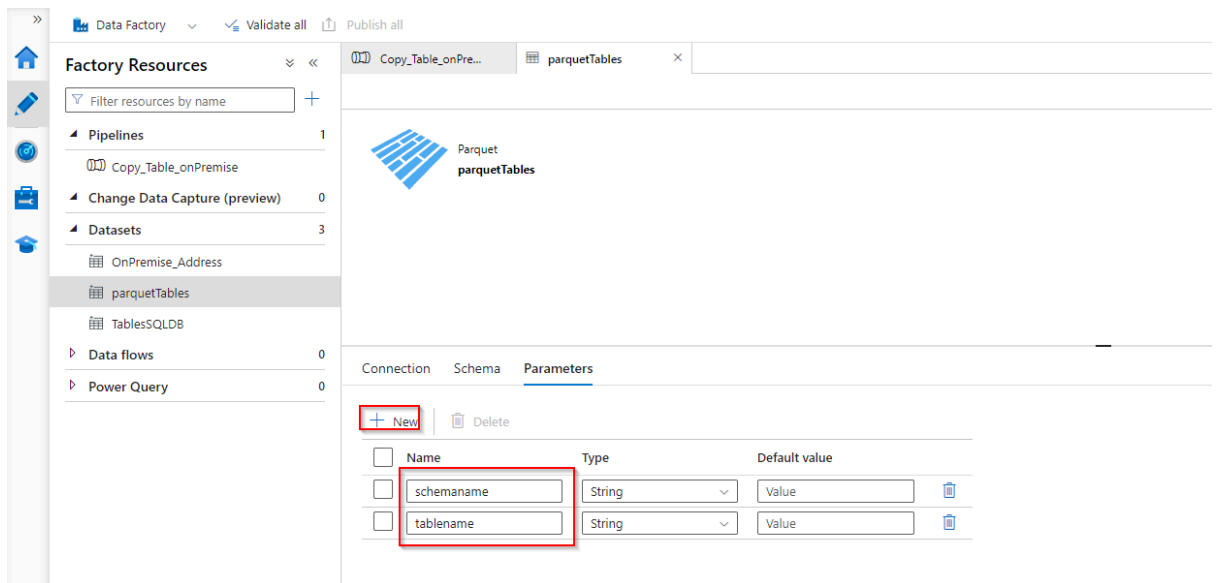
15. Click “ok” then go to datasets and select your parquet dataset, that you created and in the “file path” specified the following expressions:

Directory: `@{concat(dataset().schemaname, '/', dataset().tablename)}`

File name: `@{concat(dataset().tablename, '.parquet')}`



16. Go to Parameters and create two parameters: schemaname and tablename

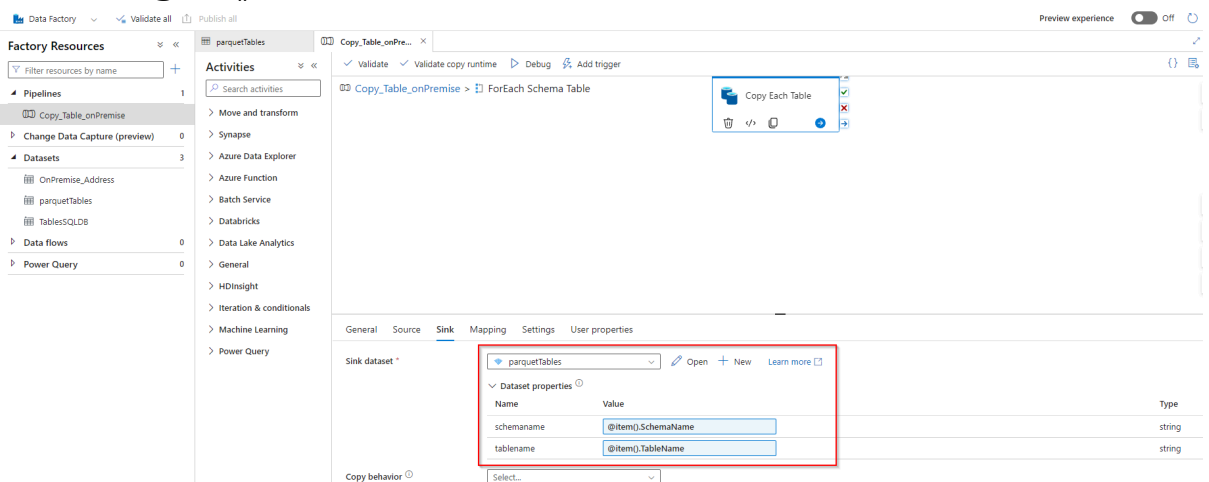


17. Go back to pipeline → copy data activity (sink section) and select your parquet dataset

18. In the dataset properties for the schemaname and tablename using following expressions:

schemaname: @item().SchemaName

tablename: @item().TableName

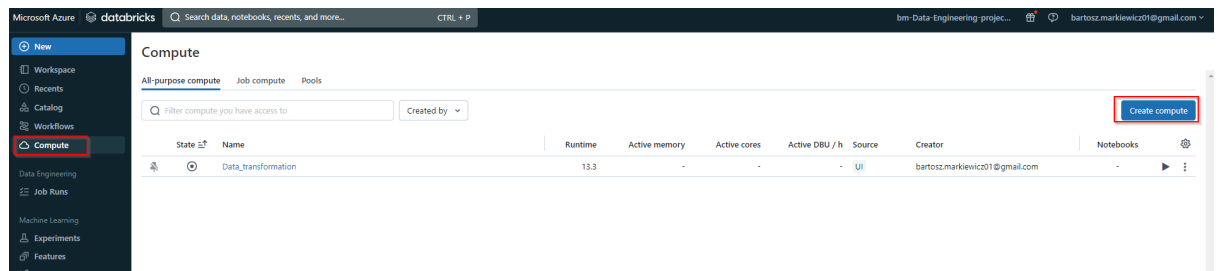


Data Transformation: Azure Databricks - setup

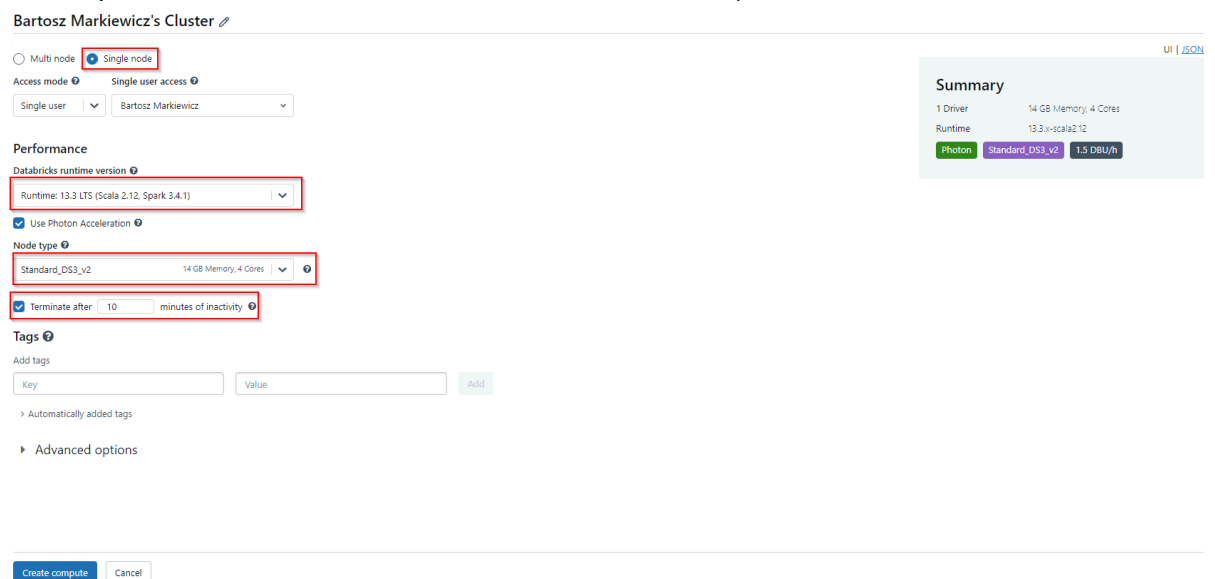
1. Firstly you need to configure the Azure Data Lake Storage Gen2, to allow mount it in databricks (if you have Azure Databricks premium you don't have to do that), link below:

<https://www.linkedin.com/pulse/how-mount-adls-gen-2-storage-account-databricks-an-any-nayak/>

2. In the Azure Databricks create new Compute



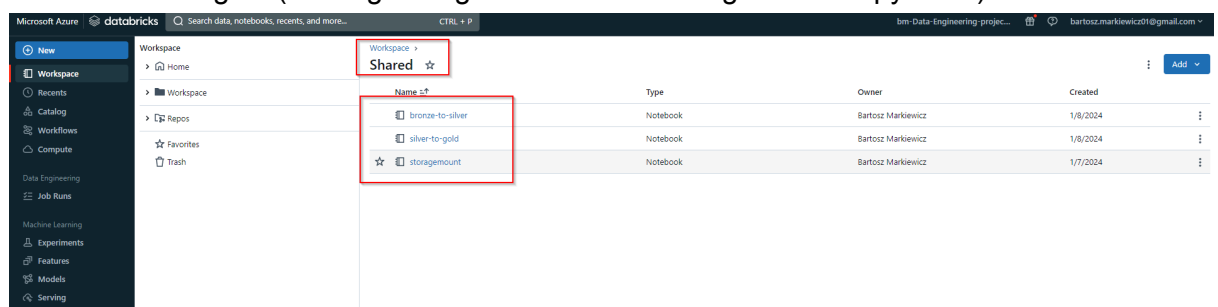
3. Set up single node, databricks runtime version, node type, terminate (it is important to set up terminate, because it will disable used machine)



4. Go to workspace then shared

5. Create three Notebooks:

- a. storagemount (in the github go to folder storagemount and copy code)
- b. bronze-to-silver (in the github go to folder bronze-to-silver and copy code)
- c. silver-to-gold (in the github go to folder silver-to-gold and copy code)



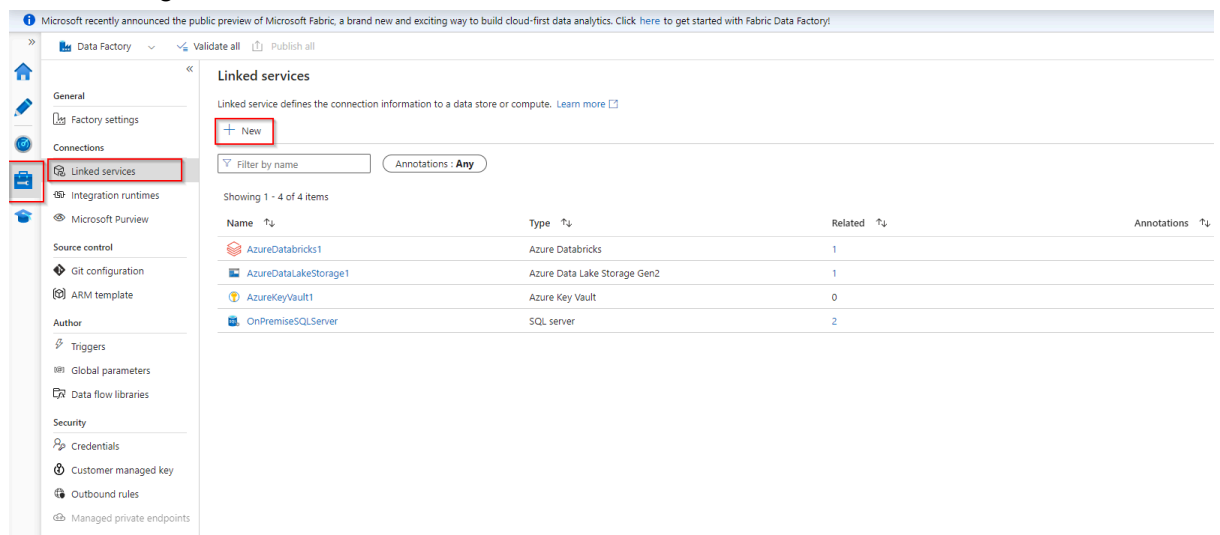
6. Notebook “storagemount”, you need to execute ones, to integrate mount Databricks with Azure Data Lake Storage Gen2
7. Description of the code you can find in the GitHub

Databricks in the ADF (azure data factory)

Configuration Linked service

You need to specify a connection in Azure data factory to azure databricks, to trigger notebooks, that have been written in the Azure databricks.

1. Go to manage → linked services and click “New”




2. Select Azure databricks


Data store Compute

Search


All Azure Compute




Azure Batch




Azure Data Lake Analytics




Azure Databricks




Azure Function




Azure HDInsight



Azure Machine Learning



Azure Machine Learning Studio (classic)



Azure Synapse Analytics (Artifacts)

Continue Cancel

3. Specify the following fields:

- Name: Your name
- Connect via Integration runtime: AutoResolveIntegrationRuntime
- Account selection method: From Azure subscription
- Azure subscription: Your subscription
- Databricks workspace: Your workspace
- Select cluster: existing interactive cluster
- Authentication type: Access token
- Access token: Token you generate in azure databricks (look second screenshot)

New linked service

 Azure Databricks [Learn more](#) 

Name *

AzureDatabricks2

Description

Connect via integration runtime * 

AutoResolveIntegrationRuntime

Account selection method *

☒ From Azure subscription ☐ Enter manually

Azure subscription * 

Azure subscription 1

Databricks workspace * 

bm-Data-Engineering-project-Databricks

Select cluster

☐ New job cluster ☒ Existing interactive cluster ☐ Existing instance pool

Databrick Workspace URL * 

4.azure.databricks.net

Authentication type *

Access Token

Access token

Azure Key Vault

Access token * 

Existing cluster ID * 

Add workspace and access token to list options

Create

Back

 Test connection

Cancel

Settings

- Workspace admin
- Identity and access
- Security
- Compute
- Development
- Notifications
- Advanced
- User
- Profile
- Preferences
- Developer**
- Linked accounts
- Notifications

Developer

Manage your development settings

Access tokens

Set up secure authentication to Databricks API using access tokens

Manage

Editor settings

General

Notebook Notifications

Controls whether browser notifications are shown for common Notebook events like when a cell is finished running.

On ☒

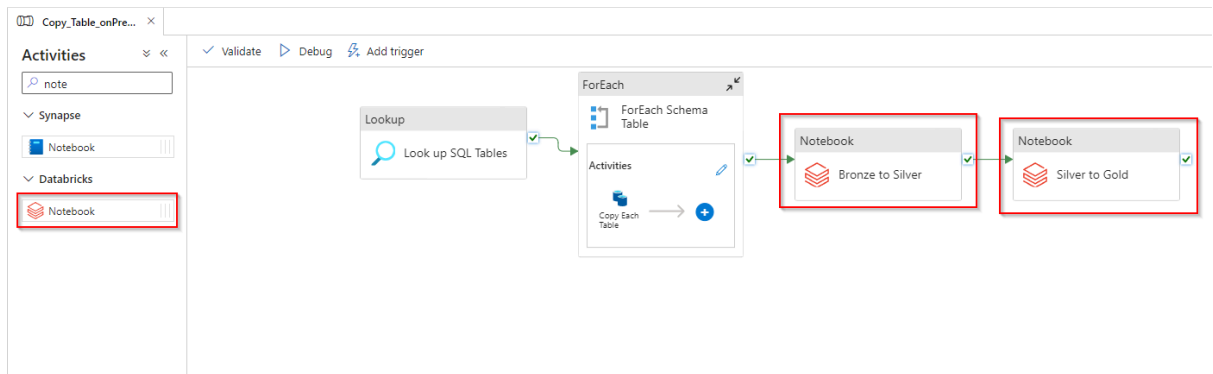
Spark tips

Enriches notebook error stack traces by displaying high-level "error hints" which explain the underlying configuration errors.

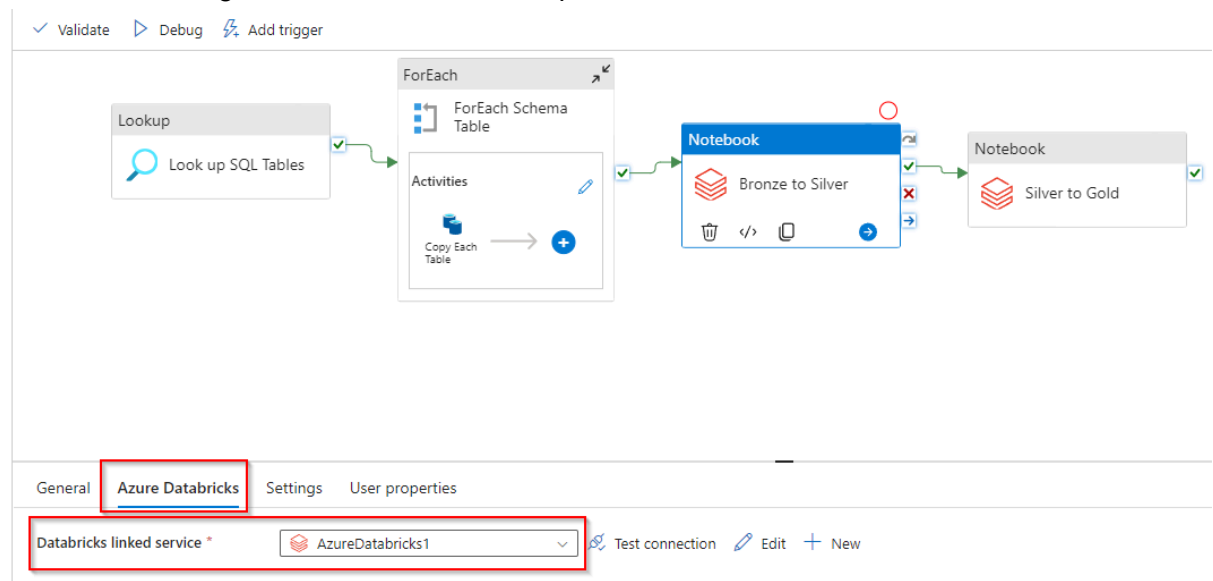
On ☒

Pipeline Configuration

1. Create new activities "notebook databricks": "bronze to silver" and "silver to gold"



2. In the Notebook go to "Azure Databricks" option and select linked service



3. In the settings option click on “Browse” and select “Shared” → notebook “bronze-to-silver”

The screenshot displays the Azure Databricks workspace interface. At the top, a workflow is visible with the following sequence of activities: a 'Lookup' activity (Look up SQL Tables), followed by a 'ForEach' activity (ForEach Schema Table) which contains a 'Copy Each Table' activity, then a 'Notebook' activity (Bronze to Silver), and finally another 'Notebook' activity (Silver to Gold). The 'Settings' tab is selected at the bottom, showing the 'Notebook path' as '/Shared/bronze-to-silver'. A red box highlights the 'Browse' button next to the path field. Below the path field, there are expandable sections for 'Base parameters' and 'Append libraries'.

Copy_Table_onPre...

Activities

note

Synapse

Notebook

Databricks

Notebook

Validate Debug Add trigger

Lookup

Look up SQL Tables

ForEach

ForEach Schema Table

Activities

Copy Each Table

Notebook

Bronze to Silver

Notebook

Silver to Gold

General Azure Databricks Settings User properties

Notebook path *

/Shared/bronze-to-silver

Browse

Open

> Base parameters

> Append libraries

started with Fabric Data Factory!

ForEach

ForEach Schema Table

Activities

Copy Each Table

✓ → Notebook

Bronze to Silver

Properties

silver

Browse

Open

Browse

Select a file or folder.

Root folder

Repos

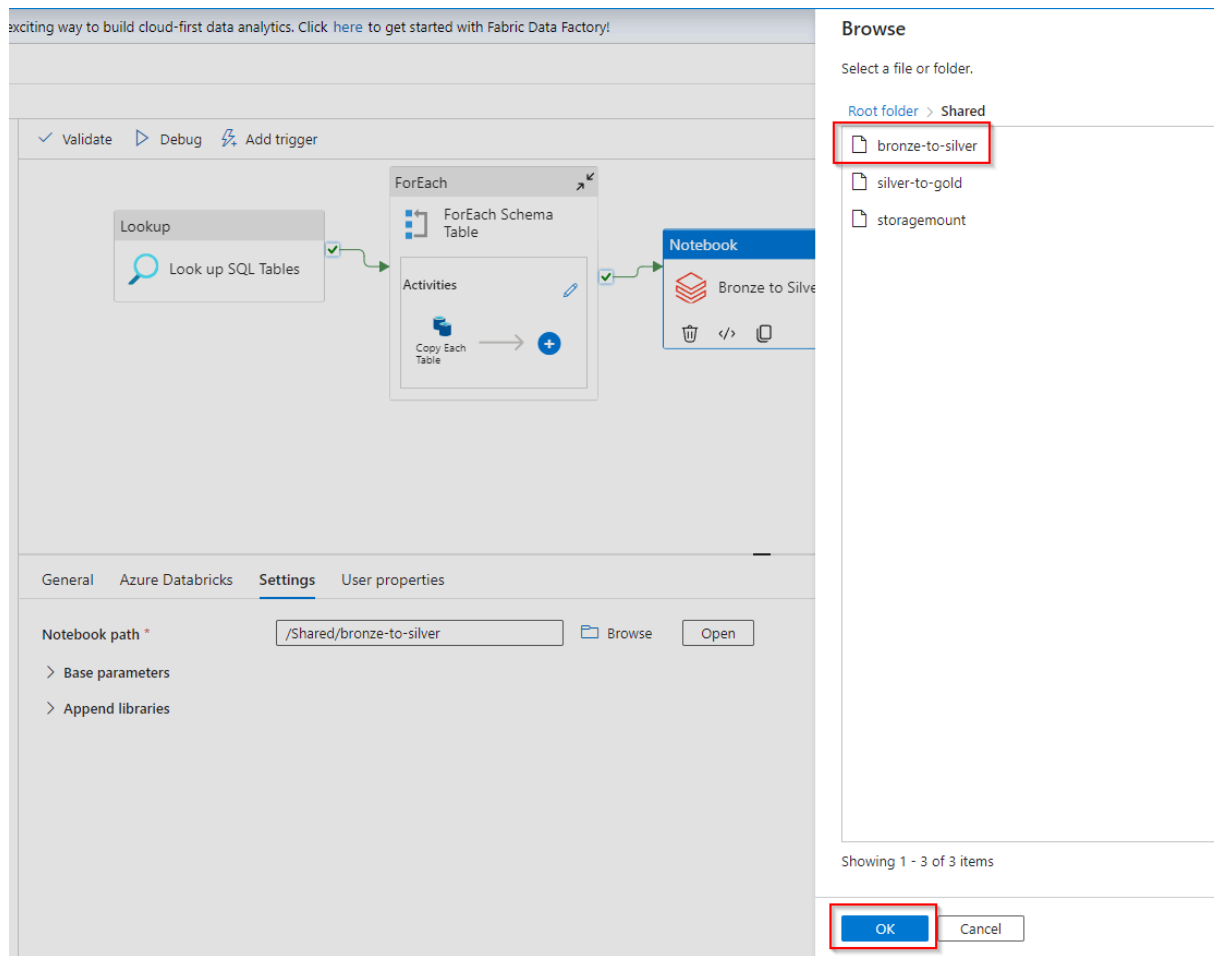
Shared

Users

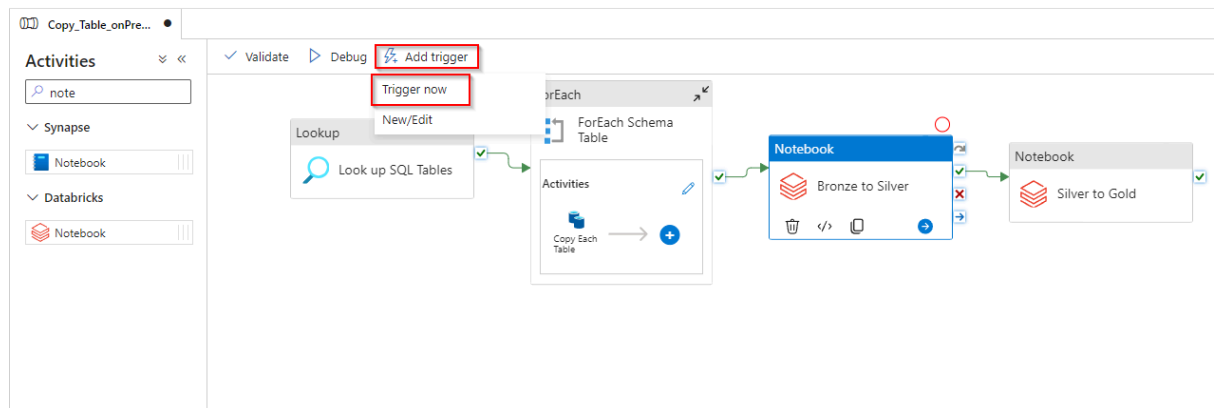
Showing 1 - 3 of 3 items

OK

Cancel



4. Do the same for the “Silver to gold”
5. Trigger the pipeline



6. Check storage account

Home > bmdatalakegen2

bmdatalakegen2 | Containers

Storage account

Search

+ Container Change access level Restore containers Refresh Delete Give feedback

Search containers by prefix

Show deleted containers

Name	Last modified	Anonymous access level	Lease state
<input type="checkbox"/> \$logs	1/4/2024, 9:47:59 PM	Private	Available
<input type="checkbox"/> bronze	1/5/2024, 11:45:53 PM	Private	Available
<input type="checkbox"/> gold	1/8/2024, 8:27:07 PM	Private	Available
<input type="checkbox"/> silver	1/8/2024, 8:27:03 PM	Private	Available

Overview

Activity log

Tags

Diagnose and solve problems

Access Control (IAM)

Data migration

Events

Storage browser

Storage Mover

Data storage

Containers

File shares

Queues

Tables

silver

Container

Search

Upload Change access level Refresh Delete Change tier Acquire lease Break lease View snapshots Create snapshot Give feedback

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: silver

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/> SalesLT						
<input type="checkbox"/> SalesLT	1/8/2024, 8:52:17 PM	Hot (Inferred)		Block blob	0 B	Available

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Access policy

Properties

Metadata

Home > bmdatalakegen2 | Containers >

silver

Container

Search

Upload Change access level Refresh Delete Change tier Acquire lease Break lease View snapshots Create snapshot Give feedback

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: silver / SalesLT

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/> [.]						
<input type="checkbox"/> Address						
<input type="checkbox"/> Customer						
<input type="checkbox"/> CustomerAddress						
<input type="checkbox"/> Product						
<input type="checkbox"/> ProductCategory						
<input type="checkbox"/> ProductDescription						
<input type="checkbox"/> ProductModel						
<input type="checkbox"/> ProductModelProductDescription						
<input type="checkbox"/> SalesOrderDetail						
<input type="checkbox"/> SalesOrderHeader						
<input type="checkbox"/> Address	1/8/2024, 8:52:17 PM	Hot (Inferred)		Block blob	0 B	Available
<input type="checkbox"/> Customer	1/8/2024, 8:52:26 PM	Hot (Inferred)		Block blob	0 B	Available
<input type="checkbox"/> CustomerAddress	1/8/2024, 8:52:29 PM	Hot (Inferred)		Block blob	0 B	Available
<input type="checkbox"/> Product	1/8/2024, 8:52:32 PM	Hot (Inferred)		Block blob	0 B	Available
<input type="checkbox"/> ProductCategory	1/8/2024, 8:52:35 PM	Hot (Inferred)		Block blob	0 B	Available
<input type="checkbox"/> ProductDescription	1/8/2024, 8:52:38 PM	Hot (Inferred)		Block blob	0 B	Available

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Access policy

Properties

Metadata

Home > bmdatalakegen2 | Containers >

silver

Container

Search

Upload Change access level Refresh Delete Change tier Acquire lease Break lease View snapshots Create snapshot Give feedback

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: silver / SalesLT / Address

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/> [.]						
<input type="checkbox"/> _delta_log						
<input type="checkbox"/> _delta_log	1/8/2024, 8:52:17 PM	Hot (Inferred)		Block blob	0 B	Available
<input type="checkbox"/> part-00000-20b19350-15b1-43af-9535-d768d898fe9.c000.snappy.parquet	1/8/2024, 8:52:19 PM	Hot (Inferred)		Block blob	34.31 KiB	Available
<input type="checkbox"/> part-00000-c49455ad-97c2-4f86-84a7-d6d3cb346490.c000.snappy.parquet	1/9/2024, 11:57:41 PM	Hot (Inferred)		Block blob	34.31 KiB	Available

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Access policy

Properties

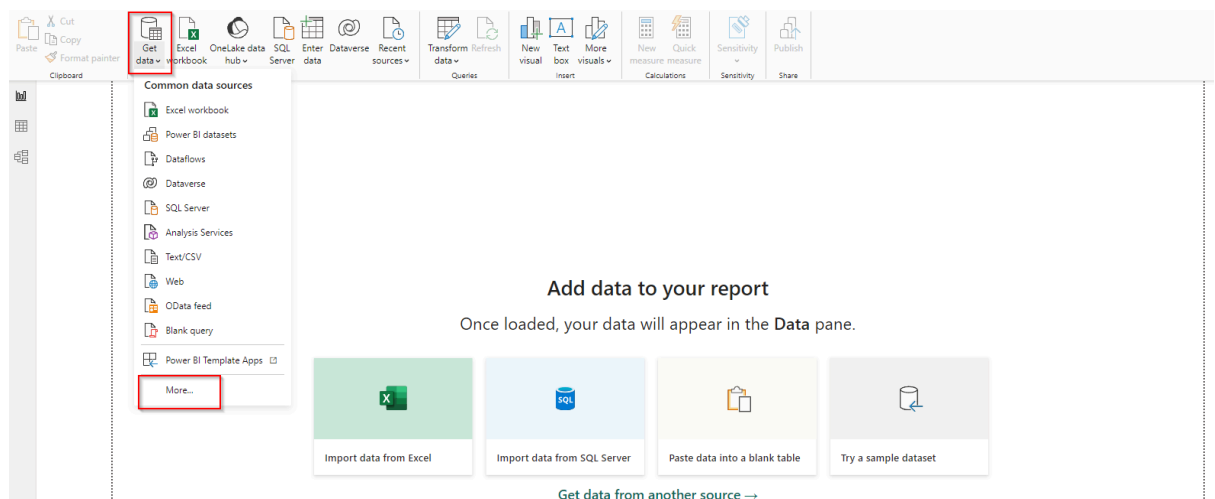
Metadata

PowerBI - load transformed data

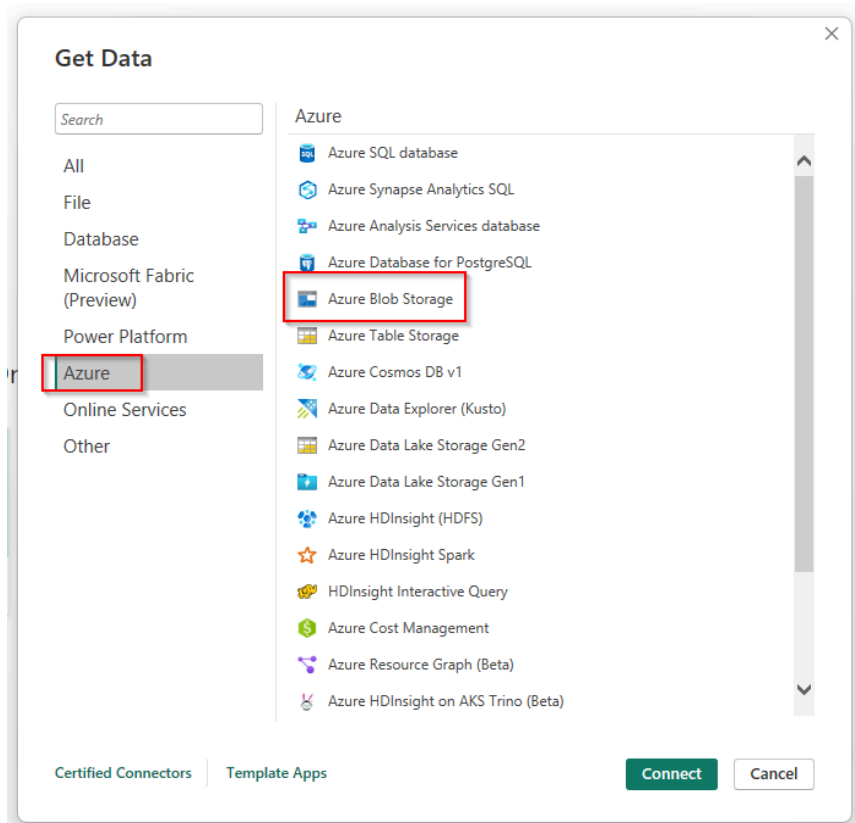
To load data to PowerBI, you can use two connectors Azure Blob storage and Azure Data Lake storage gen2

Azure Blob storage

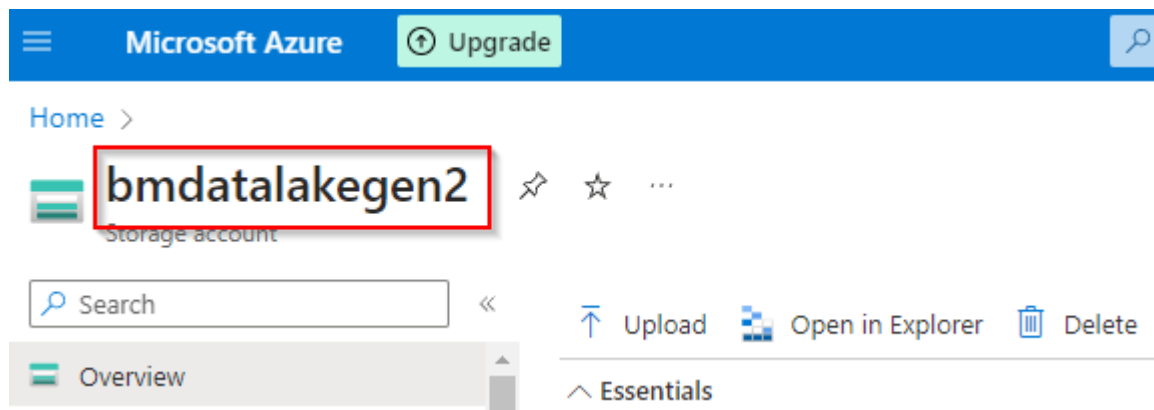
1. In the PowerBI click “Get data” then more



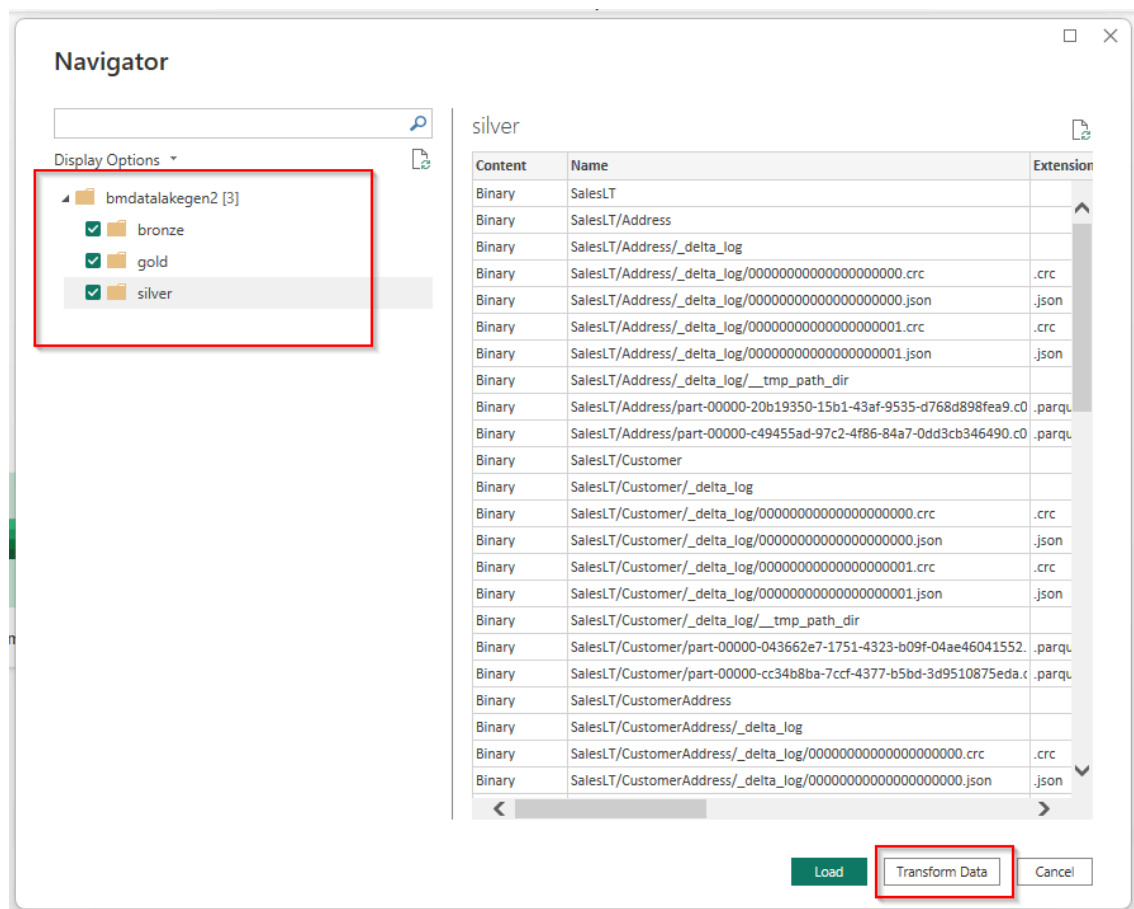
2. In the “Get data” select “Azure” and “Azure Blob Storage”



3. Type your <storage account name>, that you can find in the “Storage account” service



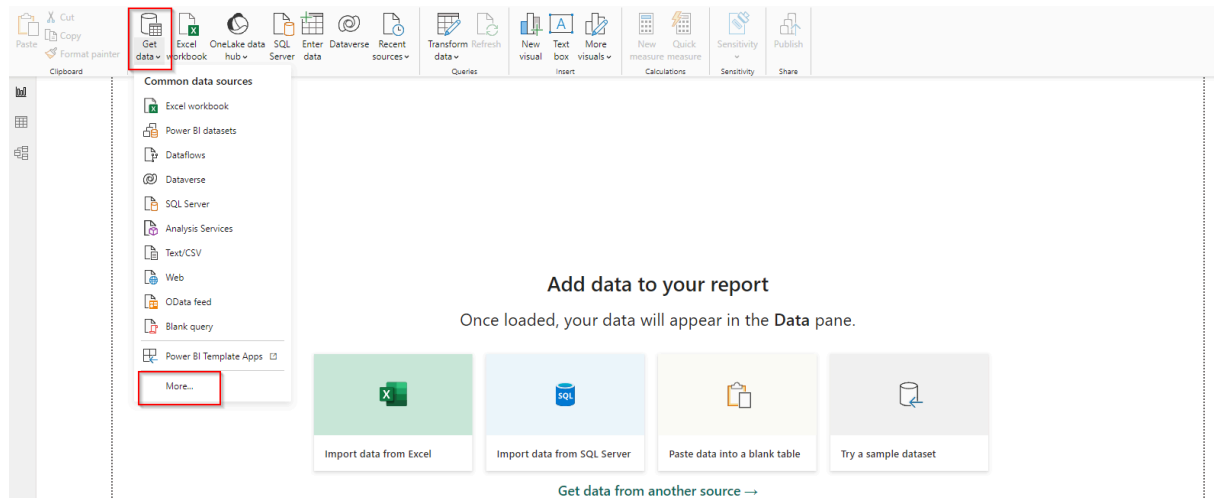
4. Select all folders and click Transform data.



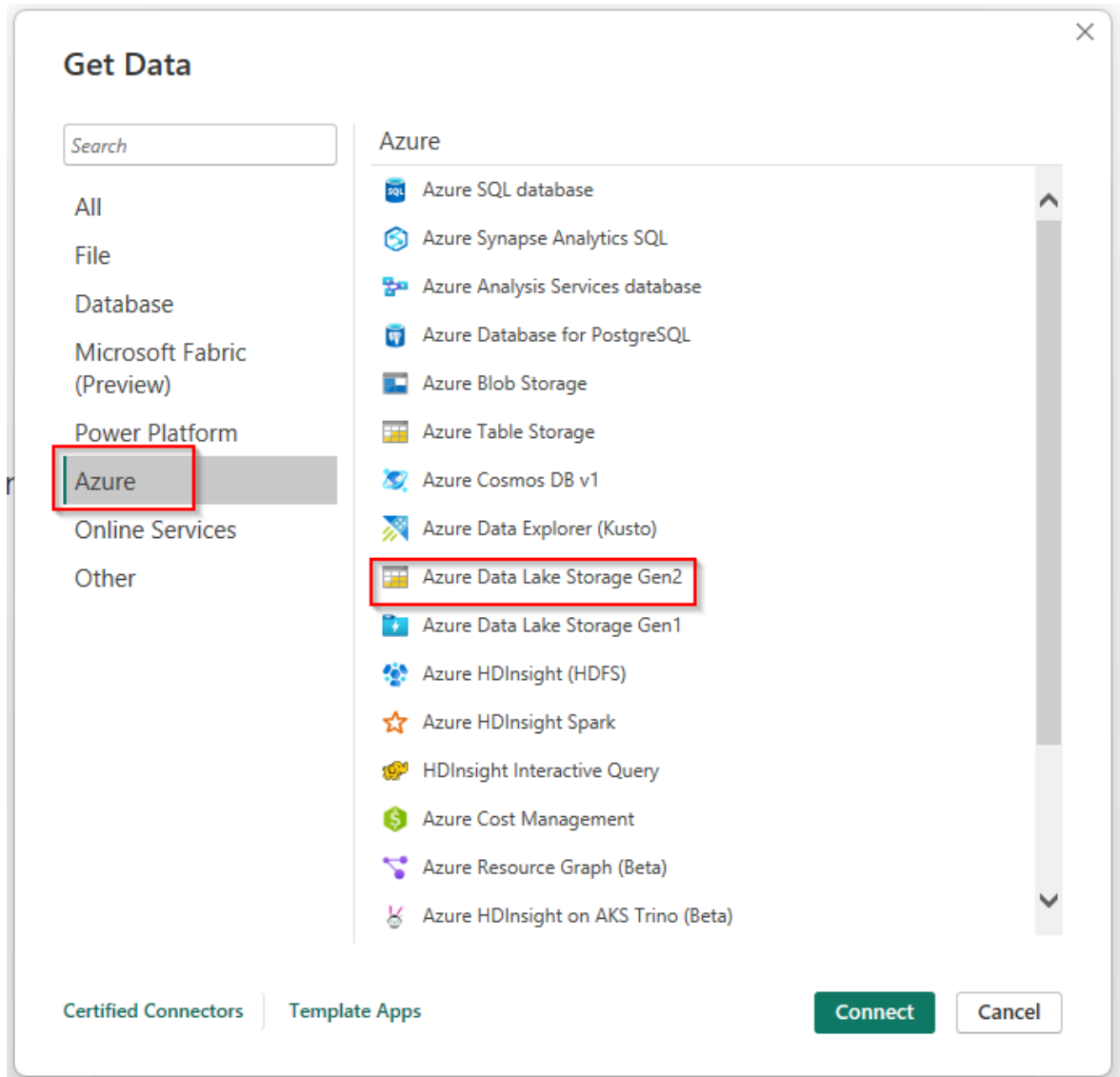
Following these steps will enable you to create the data model in Power BI by loading only the necessary tables.

Azure Data lake storage gen2

1. In the PowerBI click “Get data” then “more”



2. Then select Azure and Azure Data Lake Storage Gen2



3. Provide the url to your storage account - pattern: `https://<your storage account name>.dfs.core.windows.net/`

Azure Data Lake Storage Gen2

URL
`https://bmdatalakegen2.dfs.core.windows.net/`

Data View
☒ File System View
☐ CDM Folder View (Beta)

`https://<your storage account name>.dfs.core.windows.net/`

OK Cancel

Import data from SQL Server Paste data into a blank table Try a sample dataset

4. Click Transform Data

`https://bmdatalakegen2.dfs.core.windows.net/`

Content	Name	Extension	Date accessed	Date modified	Date created	Attributes	
Binary	Address.parquet	.parquet	null	09.01.2024 22:53:32	null	Record	htt
Binary	Customer.parquet	.parquet	null	09.01.2024 22:53:37	null	Record	htt
Binary	CustomerAddress.parquet	.parquet	null	09.01.2024 22:53:34	null	Record	htt
Binary	Product.parquet	.parquet	null	09.01.2024 22:53:34	null	Record	htt
Binary	ProductCategory.parquet	.parquet	null	09.01.2024 22:53:37	null	Record	htt
Binary	ProductDescription.parquet	.parquet	null	09.01.2024 22:53:34	null	Record	htt
Binary	ProductModel.parquet	.parquet	null	09.01.2024 22:53:34	null	Record	htt
Binary	ProductModelProductDescription.parquet	.parquet	null	09.01.2024 22:53:34	null	Record	htt
Binary	SalesOrderDetail.parquet	.parquet	null	09.01.2024 22:53:33	null	Record	htt
Binary	SalesOrderHeader.parquet	.parquet	null	09.01.2024 22:53:34	null	Record	htt
Binary	000000000000000000000000.crc	.crc	null	08.01.2024 20:12:39	null	Record	htt
Binary	000000000000000000000000.json	.json	null	08.01.2024 20:12:35	null	Record	htt
Binary	000000000000000000000001.crc	.crc	null	08.01.2024 20:14:57	null	Record	htt
Binary	000000000000000000000001.json	.json	null	08.01.2024 20:14:56	null	Record	htt
Binary	000000000000000000000002.crc	.crc	null	08.01.2024 20:16:54	null	Record	htt
Binary	000000000000000000000002.json	.json	null	08.01.2024 20:16:54	null	Record	htt
Binary	000000000000000000000003.crc	.crc	null	09.01.2024 23:02:01	null	Record	htt
Binary	000000000000000000000003.json	.json	null	09.01.2024 23:02:00	null	Record	htt
Binary	part-00000-920840d8-72d5-4c99-93f3-1a1da3c03843.c...	.parquet	null	08.01.2024 20:16:53	null	Record	htt
Binary	part-00000-96fc53c8-dfb8-4d3c-a983-a3283e543799.c...	.parquet	null	08.01.2024 20:14:56	null	Record	htt

The data in the preview has been truncated due to size limits.

Combine Load **Transform Data** Cancel

Following these steps will enable you to create the data model in Power BI by loading only the necessary tables.

Glossary

Azure Blob Storage

Azure Blob Storage is a highly scalable, durable, and secure object storage service that stores unstructured data objects, such as text, images, and videos. It is a popular choice for storing large datasets that need to be accessed frequently. Blob Storage supports several different data formats, including Avro, Parquet, and CSV.

Azure Data Lake Storage Gen2

Azure Data Lake Storage Gen2 is a high-performance, secure, and scalable data lake storage service that stores large volumes of structured, semi-structured, and unstructured data. It is designed to handle petabytes of data and can be used for a variety of purposes, including data ingestion, data processing, and data analytics. Data Lake Storage Gen2 supports a variety of data formats, including Parquet, Avro, and JSON.

Azure Data Lake Storage Gen

Azure Data Lake Storage Gen was a previous version of Azure Data Lake Storage Gen2. It was a less scalable and less secure than Gen2, but it was still a popular choice for storing large datasets. Gen is no longer supported, and all new deployments should use Gen2.

Difference between Azure Data Lake Storage Gen vs Azure Data Lake Storage Gen2

Here is a table that summarizes the key differences between Azure Data Lake Storage Gen and Azure Data Lake Storage Gen2:

Feature	Azure Data Lake Storage Gen	Azure Data Lake Storage Gen2
Scalability	Less scalable	Highly scalable
Security	Less secure	More secure

Features	Fewer features	More features
Support	No longer supported	Supported

Parquet

Parquet is a columnar data storage format that is optimized for analytical workloads. It is commonly used in data lakes and data warehouses. Parquet files are compressed and have a hierarchical file structure that makes it efficient to read and write data.

Blob

A blob is a general-purpose data storage object that can store unstructured data, such as text, images, and videos. Blobs are typically stored in a hierarchical file system. Blob storage is a scalable and durable storage solution that is commonly used for storing large volumes of data.

Avro

Avro is a data serialization format that is designed for efficiency and flexibility. It is a self-describing format, which means that the schema of the data is stored in the file itself. Avro is commonly used for storing structured data in data lakes and data warehouses.

Delta Format

Delta Lake is a storage layer that sits on top of Apache Spark that provides a set of capabilities for managing and analyzing large datasets. Delta Lake stores data in Parquet files and uses a version control system to keep track of changes to the data. Delta Lake also provides a number of features for managing and analyzing data, such as data lineage and ACID transactions.

Total Cost of the project: 1,04 EURO