😊 Azure Data Factory

Azure Data Factory is a cloud-based data integration service provided by Microsoft Azure. It allows you to create, schedule, and manage data pipelines that move data between various supported data stores and processing systems. Here's an overview of its key features and components:

1. **Data Pipelines:** Azure Data Factory enables you to create data pipelines that orchestrate and automate the movement and transformation of data from various sources to destinations. These pipelines can include activities like data ingestion, data transformation, and data loading.
2. **Integration with Various Data Sources:** It supports integration with a wide range of data sources, including Azure Blob Storage, Azure SQL Database, Azure Data Lake Storage, Azure Synapse Analytics (formerly SQL Data Warehouse), on-premises data sources, and many others.
3. **Data Transformation:** Azure Data Factory allows you to transform data using various built-in activities such as mapping data flows, SQL transformations, and custom code activities. Data flows provide a visual interface for building data transformation logic using a drag-and-drop approach.
4. **Monitoring and Management:** It provides monitoring and management capabilities through Azure Monitor and Azure Data Factory's monitoring dashboard. You can monitor pipeline runs, activity runs, and trigger events, and set up alerts for monitoring pipeline health and performance.
5. **Integration with Other Azure Services:** Azure Data Factory integrates with other Azure services such as Azure Databricks, Azure Machine Learning, Azure Functions, and more, enabling you to build end-to-end data integration and analytics solutions.
6. **Scalability and Reliability:** It offers scalability and reliability features such as auto-scaling, fault tolerance, and retry policies to ensure that your data integration processes can handle varying workloads and maintain high availability.
7. **Security:** Azure Data Factory provides robust security features including encryption of data in transit and at rest, role-based access control (RBAC), integration with Azure Active Directory for authentication, and network security controls.

😄 Use cases of Azure data factory:

Azure Data Factory is a versatile tool with a wide range of use cases across industries. Here are some common scenarios where organizations leverage Azure Data Factory:

1. **Data Warehousing:** Companies often use Azure Data Factory to populate and manage data warehouses. It can extract data from various sources, transform it to fit the schema of the data warehouse, and load it into Azure Synapse Analytics or other data warehousing solutions.
2. **ETL (Extract, Transform, Load) Processes:** Azure Data Factory is commonly used for ETL processes, where data needs to be extracted from multiple sources, transformed according to business rules, and loaded into a target destination. This could include

aggregating data from sales systems, CRM databases, and web logs, then loading it into a data lake or data warehouse.

3. **Real-time Data Ingestion:** Organizations dealing with real-time data streams, such as IoT sensor data, social media feeds, or clickstream data, can use Azure Data Factory to ingest, process, and analyze streaming data in near real-time. It can connect to Azure Event Hubs, Azure IoT Hub, or other streaming sources for continuous data ingestion.
4. **Data Migration:** Azure Data Factory simplifies the process of migrating data from on-premises data sources to the cloud or between different cloud platforms. It supports various data migration scenarios, including database migration, file migration, and big data migration.
5. **Data Integration for Analytics:** Businesses often use Azure Data Factory to integrate data from multiple sources for analytics purposes. It can combine structured and unstructured data from sources such as databases, files, and cloud services, and transform it into formats suitable for analysis using tools like Azure Synapse Analytics, Azure Databricks, or Power BI.
6. **Data Archiving and Backup:** Azure Data Factory can automate the process of archiving and backing up data to ensure data durability and compliance. It can schedule regular backups of on-premises or cloud-based data stores to Azure Blob Storage, Azure Data Lake Storage, or Azure SQL Database.
7. **Machine Learning Pipelines:** Organizations leveraging machine learning for predictive analytics or AI applications can use Azure Data Factory to build and manage end-to-end machine learning pipelines. It can orchestrate the flow of data between data storage, data preprocessing, model training, and model deployment stages.
8. **Hybrid Data Integration:** Azure Data Factory supports hybrid data integration scenarios where data resides both in the cloud and on-premises. It can securely connect to on-premises data sources using self-hosted integration runtimes, enabling seamless integration between cloud-based and on-premises data environments.

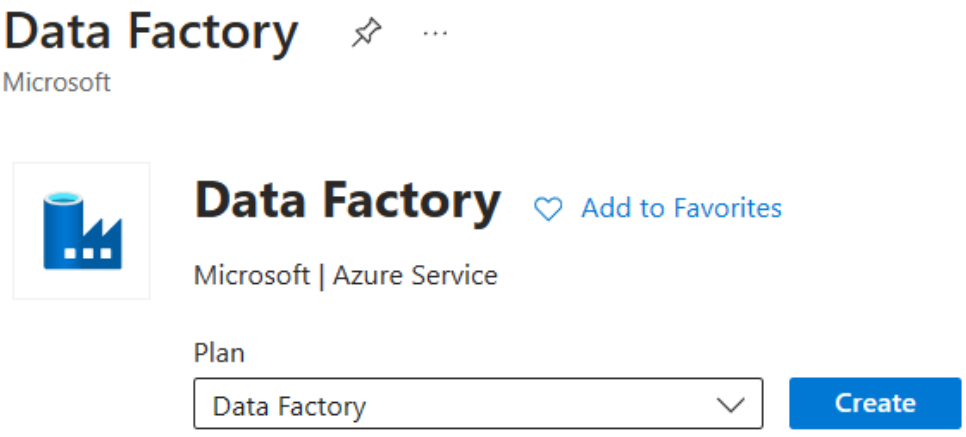
In this walkthrough, we're setting up Azure Data Factory to orchestrate the movement and transformation of data between various sources and destinations. The end goal is to establish a data integration pipeline that extracts data from a CSV file stored in Azure Data Lake Storage Gen2, transforms it, and loads it into a table in Azure Synapse Analytics. This process allows organizations to automate data workflows, ensuring data consistency, efficiency, and accessibility for analytics and decision-making purposes.

😄 To begin with the Lab:

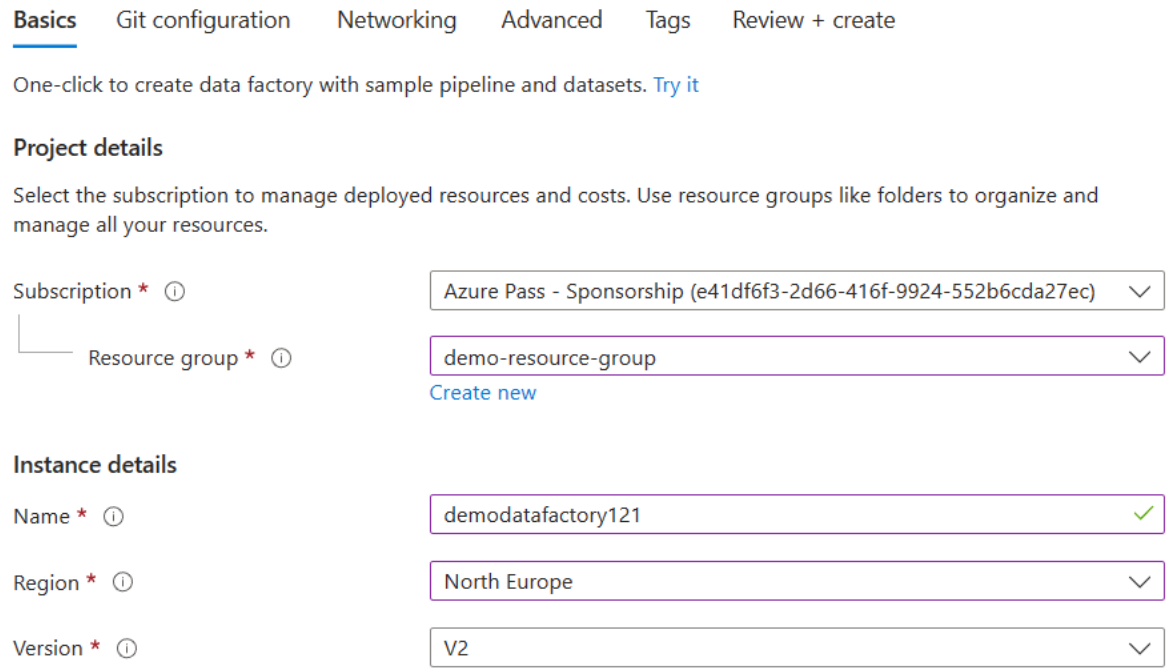
1. There are some prerequisites for this lab and they are, that you should have resources from previous labs.
2. You should have an external storage account created separately and it should have two containers one for CSV and the other one for Parquet, then the other storage account should be created while creating your Azure Synapse Workspace.
3. Then you should also have an SQL Database on which sample data should be loaded while creating it.

Name	Type	Location
datalake121	Storage account	North Europe
dataserver121	SQL server	North Europe
datasynapse1234	Synapse workspace	North Europe
demodb (dataserver121/demodb)	SQL database	North Europe
demopooldb (datasynapse1234/demopooldb)	Dedicated SQL pool	North Europe
sqlstorage1010	Storage account	North Europe

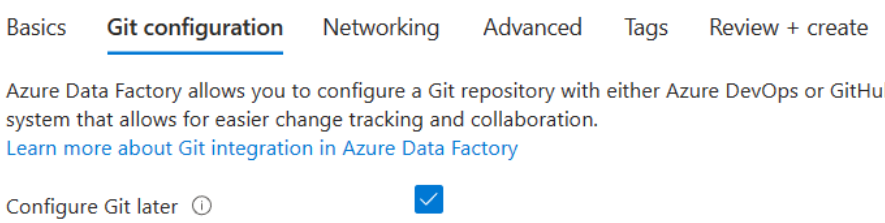
4. Now you are going to create Azure Data Factory. For that go to create resources section and search for it. Then choose this service accordingly.



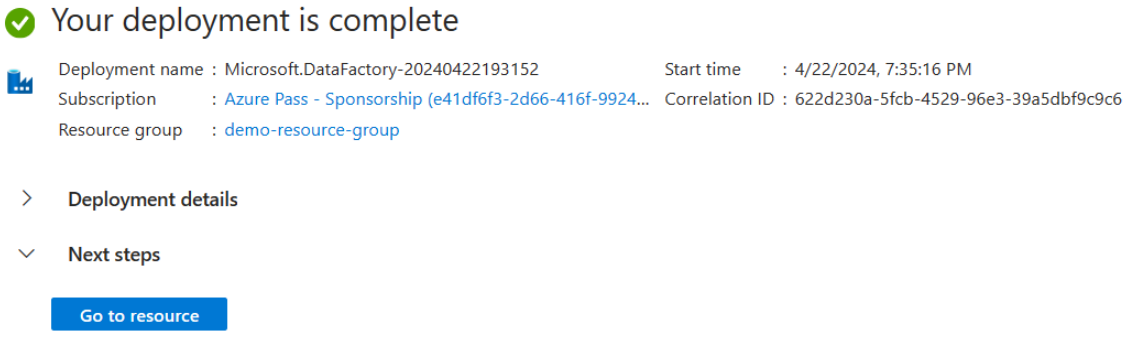
5. After that you have to choose your resource group where your all resources reside. Then give a unique name to Azure data factory. After that choose your location. Then click on next.



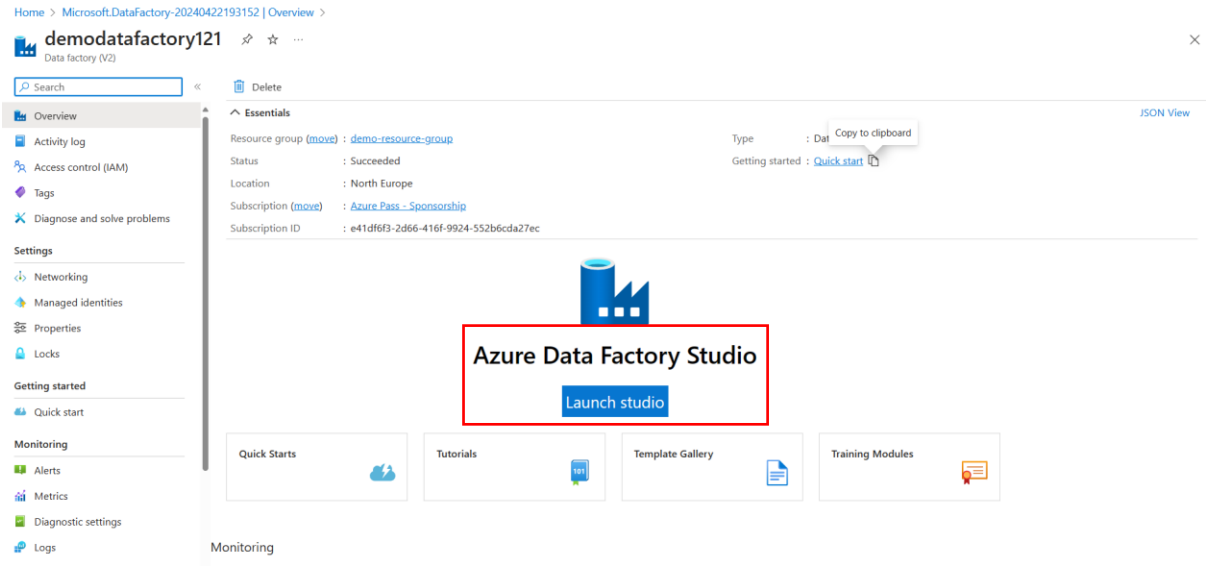
6. On the next page click on configure git later on and move directly to review page and create your resources.



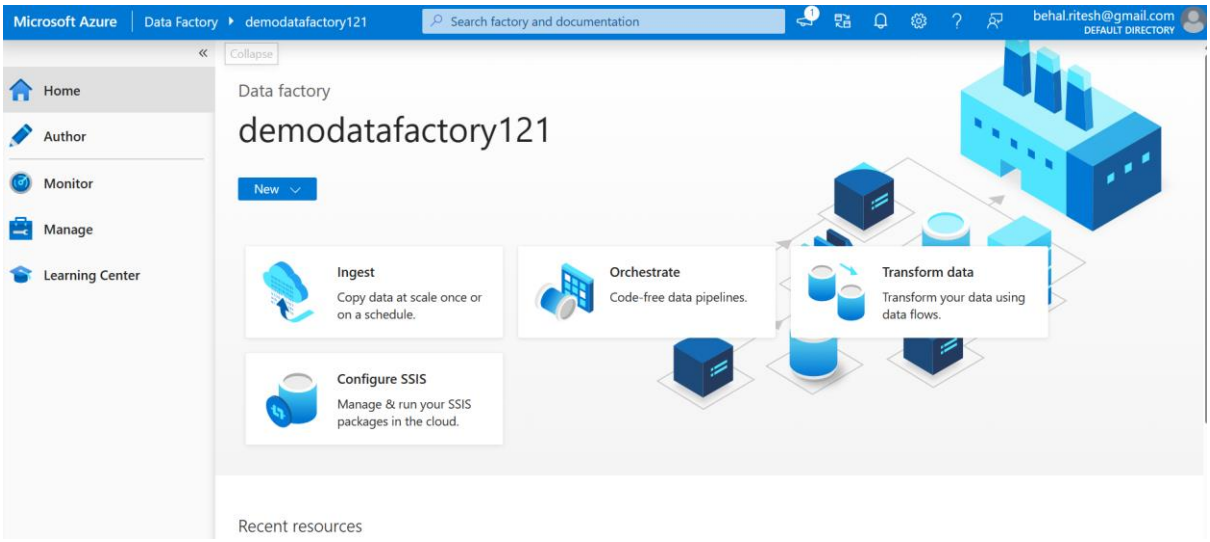
7. Now you have to wait for some time and wait until the deployment gets complete. After that click on go to resources.



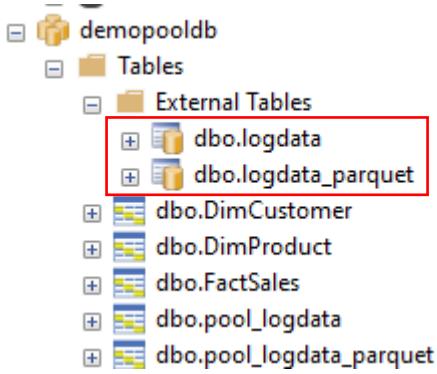
8. From its dashboard click on launch studio.



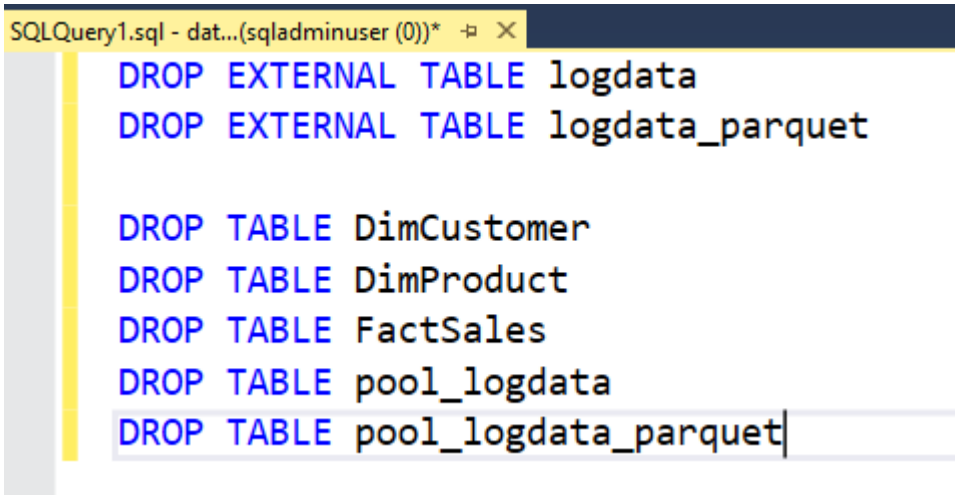
9. This will give us an entire interface in which we can start working with our pipelines, activities, our data sets, link services, etcetera.
10. Below you can see that it is similar to Azure Synapse Studio.



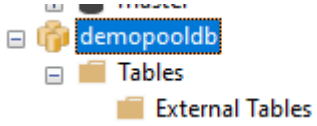
11. Now we need to get into SSMS (SQL Server Management Studio). Connect your SQL Database and Synapse Workspace.
12. Here you will expand your Pool db tables. Then you will find your external tables and normal tables which you have created while working with synapse.
13. Or if you have already delete all your tables and starting again then you just have to create synapse workspace, SQL Database and pool db in your synapse workspace. Along with that you should have your storage account in place in which you should be having two containers one for csv and other for parquet with respective files in it.



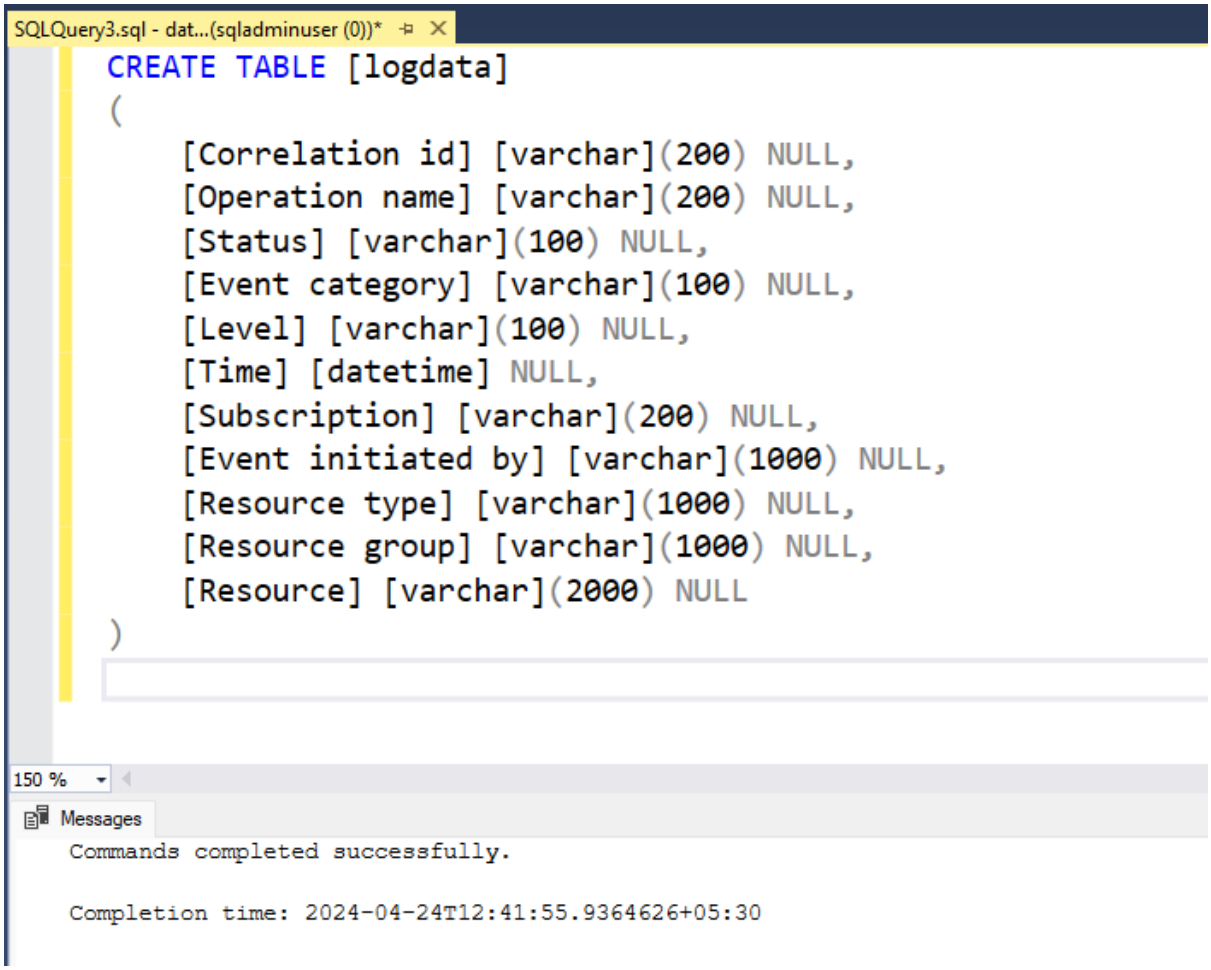
14. Now we are going to Drop these tables. We did this so that we could have a clean slate for ourselves.



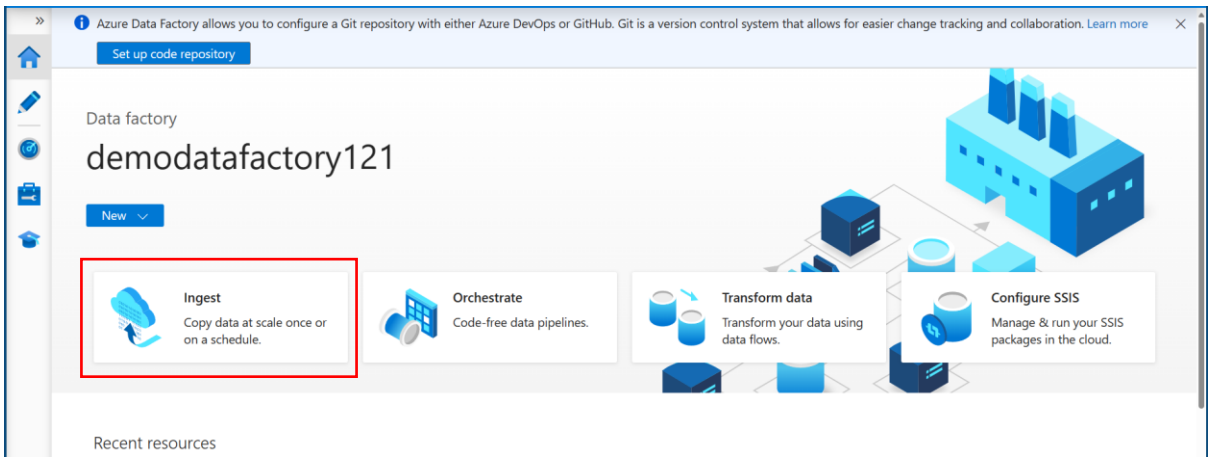
15. Now you can see that we have cleaned all of the tables.



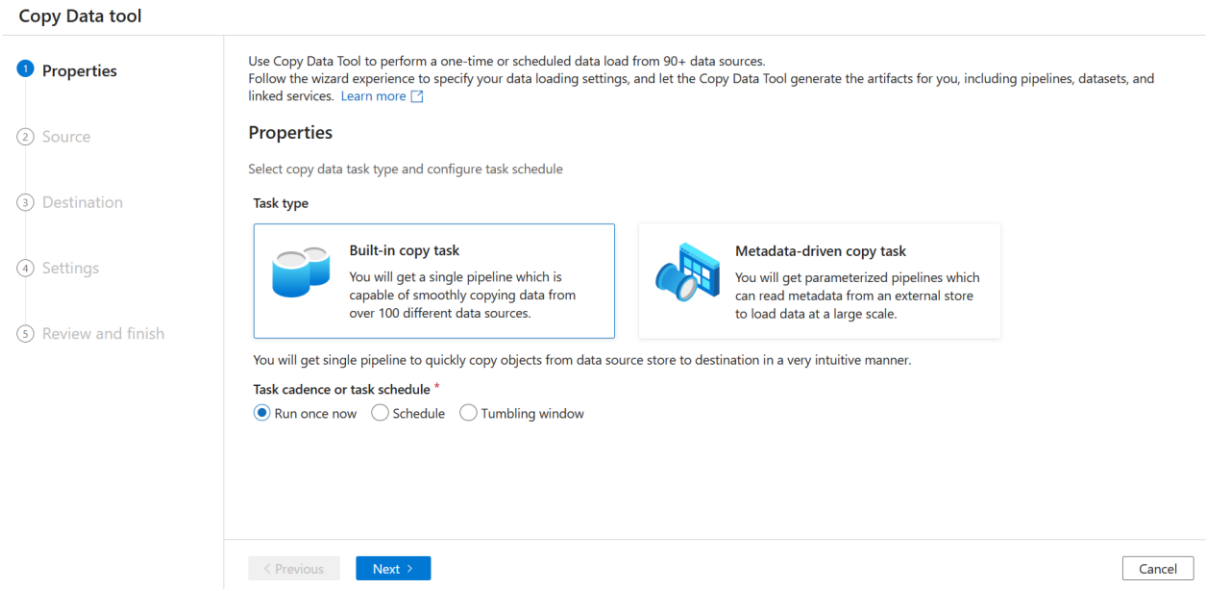
16. Now we are going to create a new table in our demo pool db. For that right click on your pool db and start a new query.
17. Then you have to paste the statement here in SSMS. There you can see that you have created a new table.



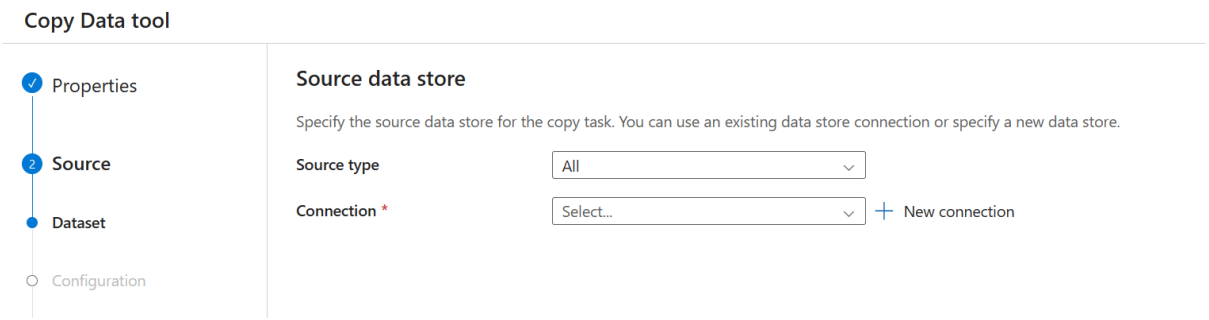
18. Now you have to navigate to Azure Data Factory. From the home section you have to click on Ingest.



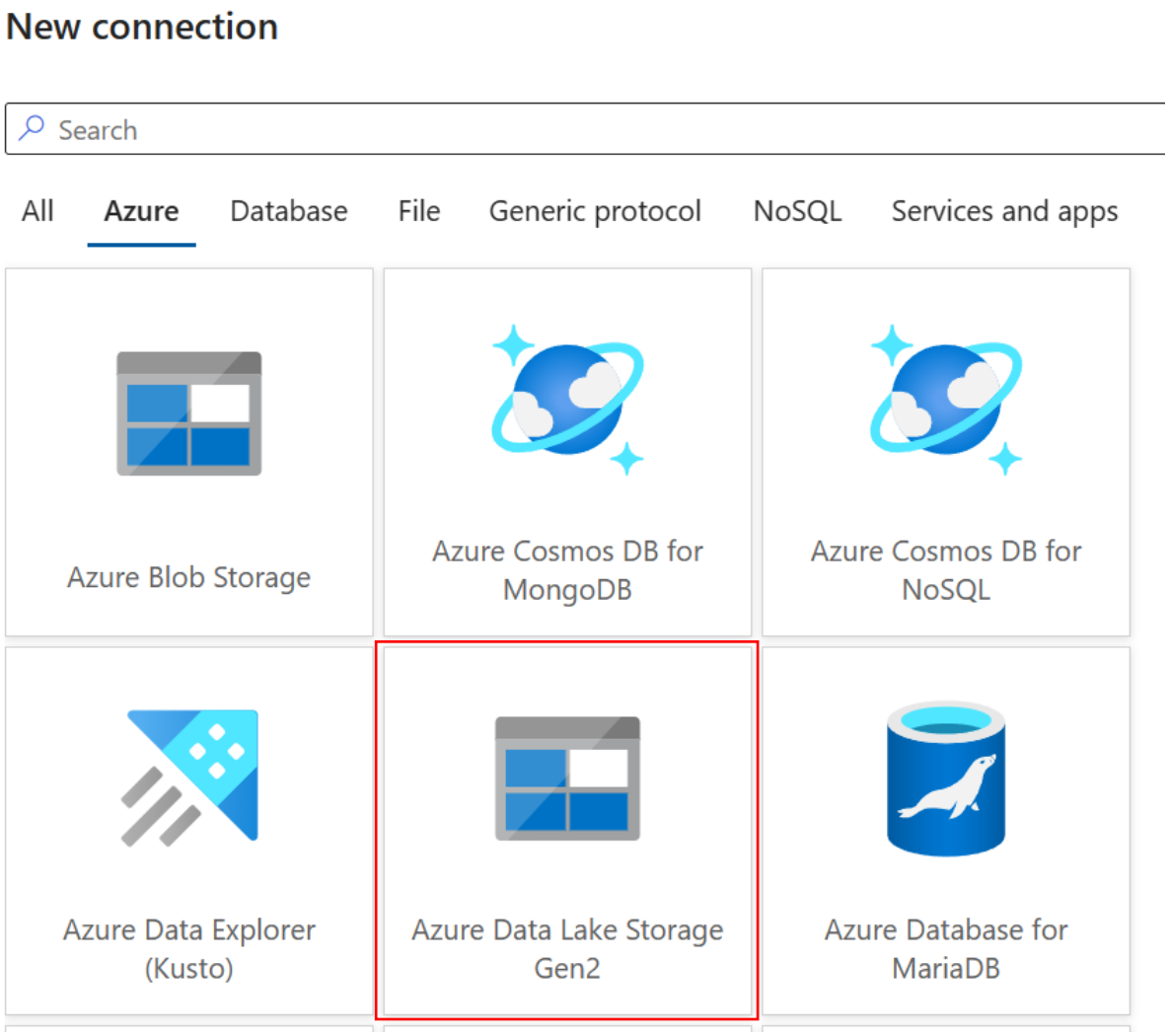
19. Here we are going to start our build-in copy task which we have seen in Synapse also. Just click on next.



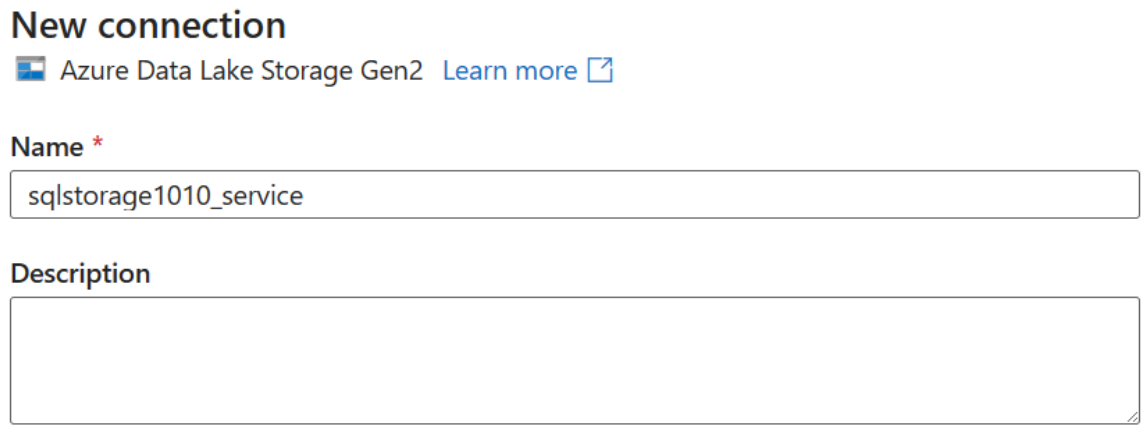
20. After that you have to choose New connection.



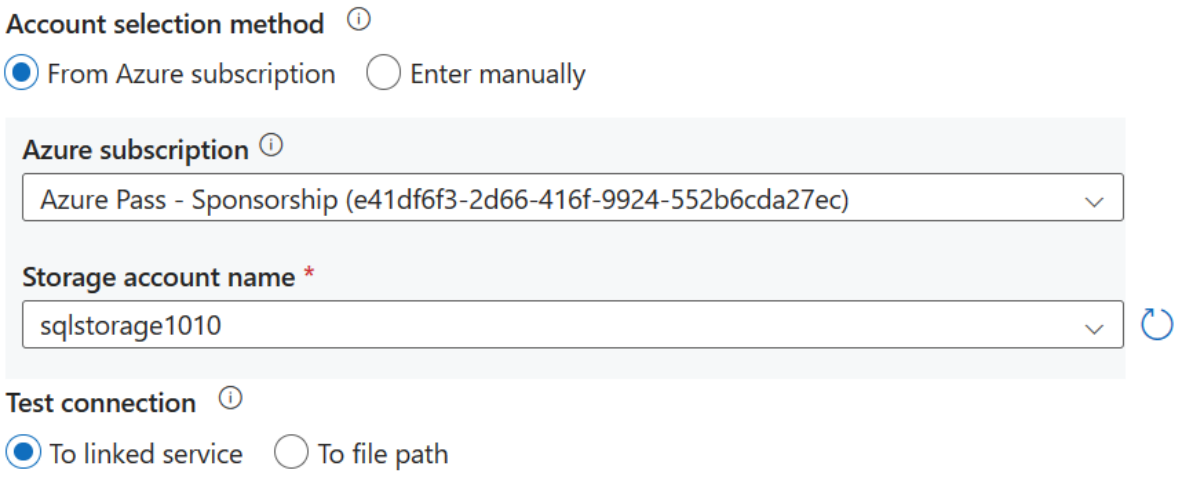
21. Then choose Azure after that choose Azure data lake storage Gen2. Select it and click on continue.



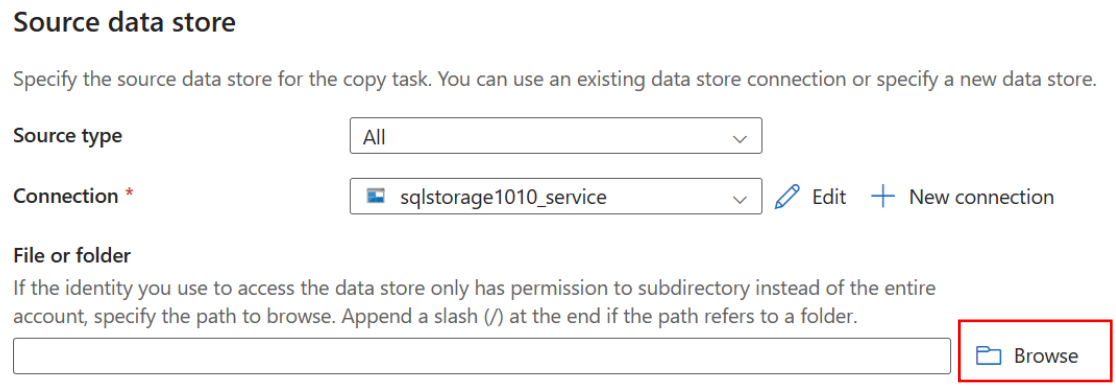
22. Then you have to give it a name, you can copy your storage account name in which your containers are.



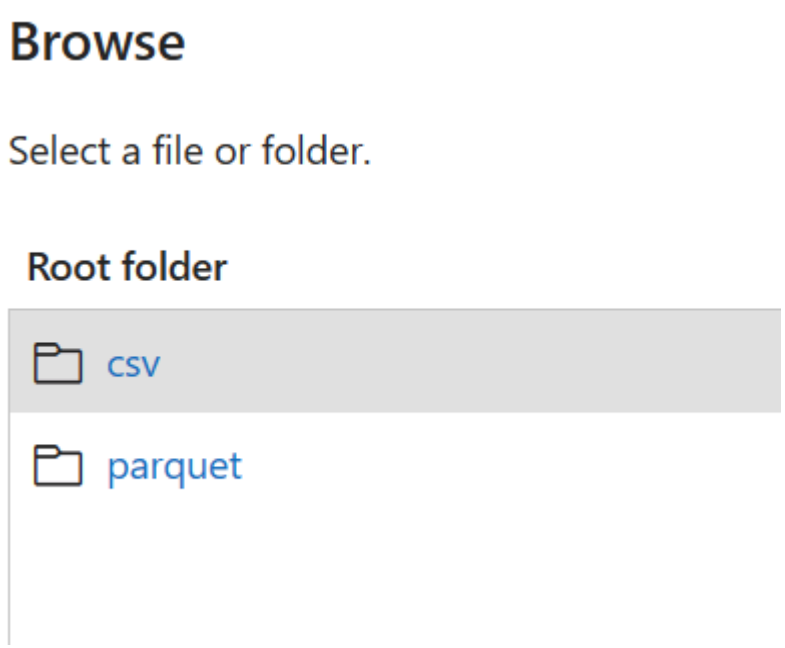
23. After that choose your subscription and choose your storage account in which you have your contains. Test connection should be linked service. Then hit on create.



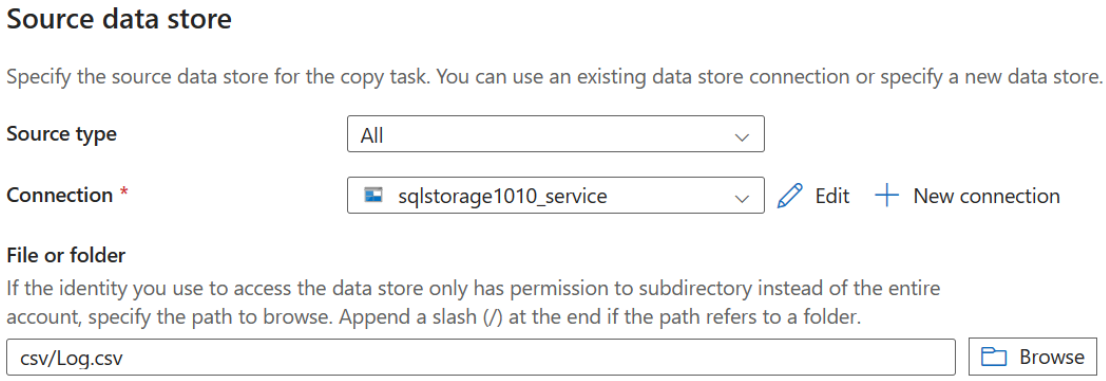
24. Once you are connected then you have to browse for you container in which you have CSV file.



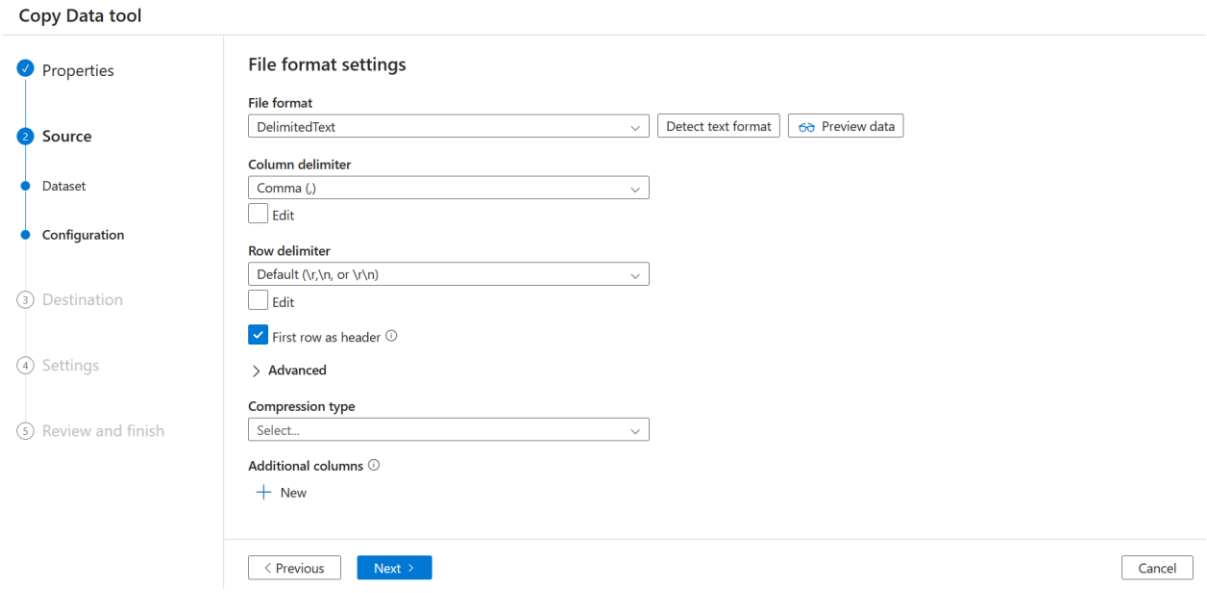
25. Choose your CSV container and go inside of it then choose your Log.csv file.



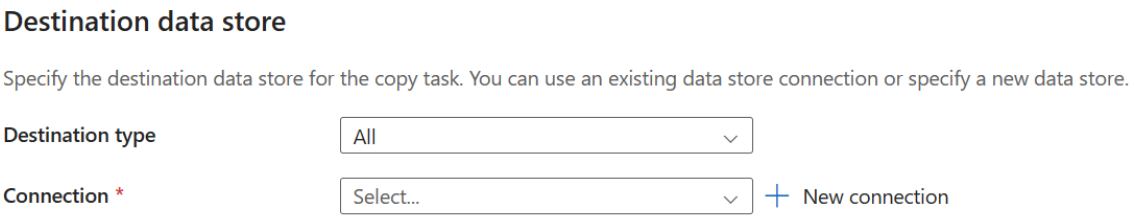
26. Below you can see that we have our file in place.



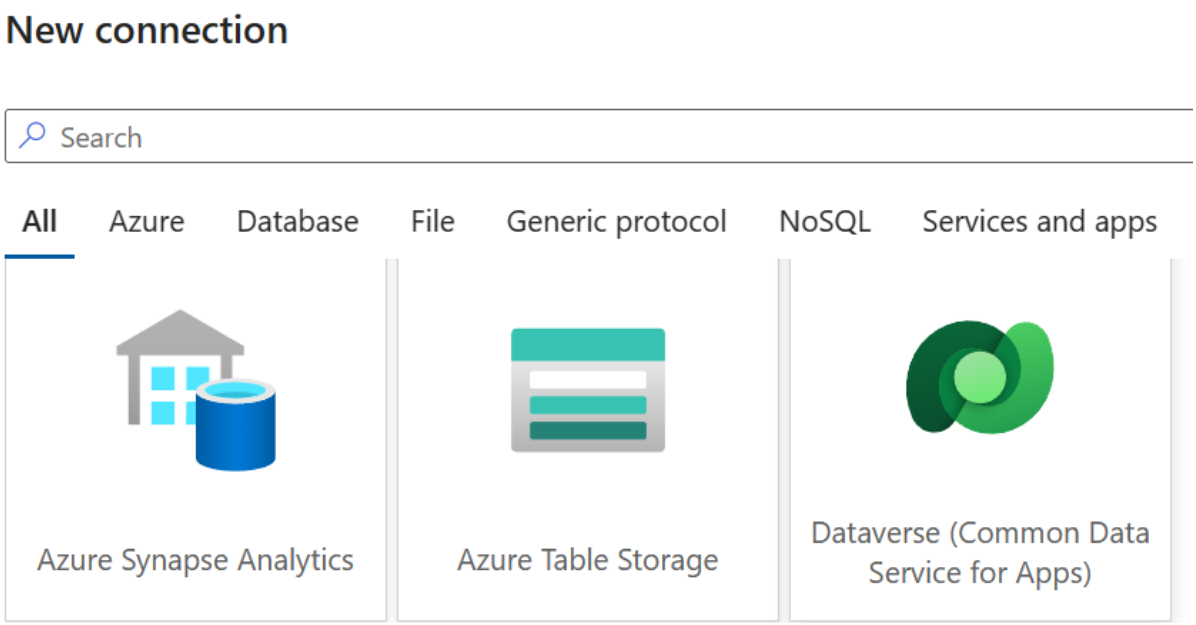
27. Then click on the next button here you will see that the wizard has identified the underlying format for our CSV file. You can also preview your data to check whether the tool has your proper data or not. After that click on next.



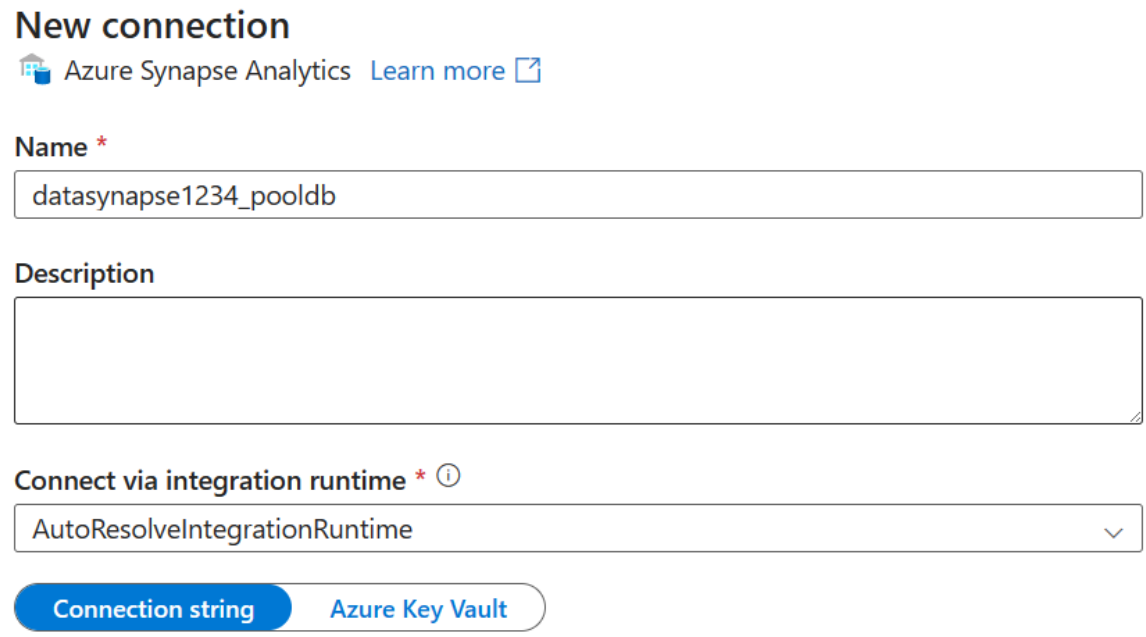
28. Now for the destination you have to click on the new connection again and this time you will choose your Synapse Workspace.



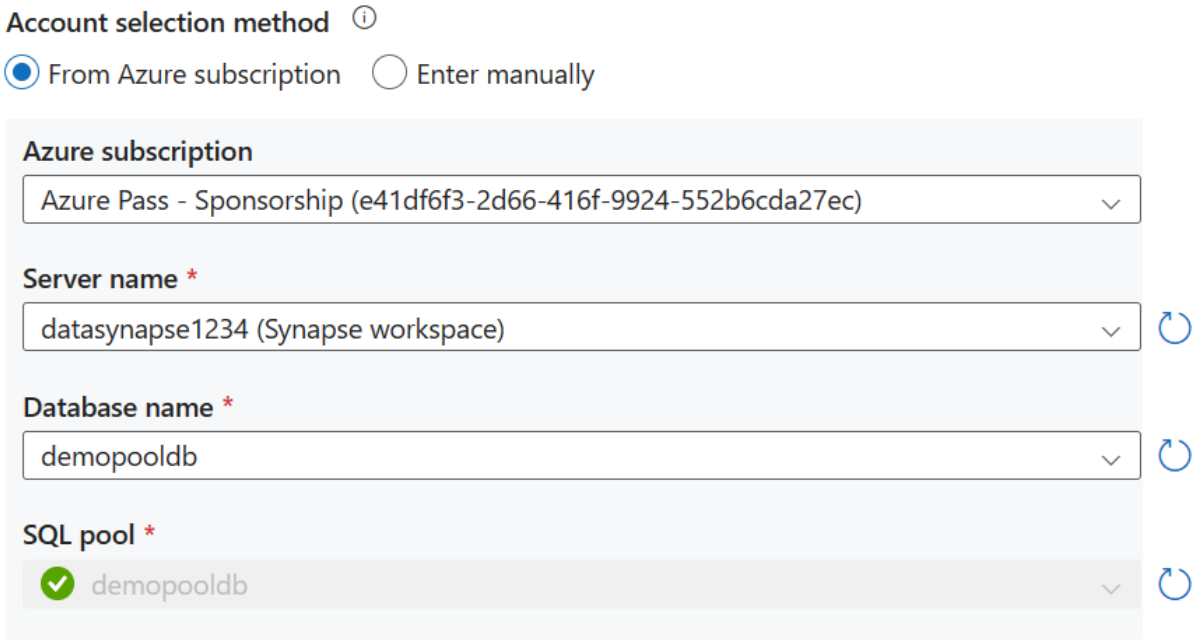
29. Here in all section you have to scroll down and find Azure Synapse Analytics



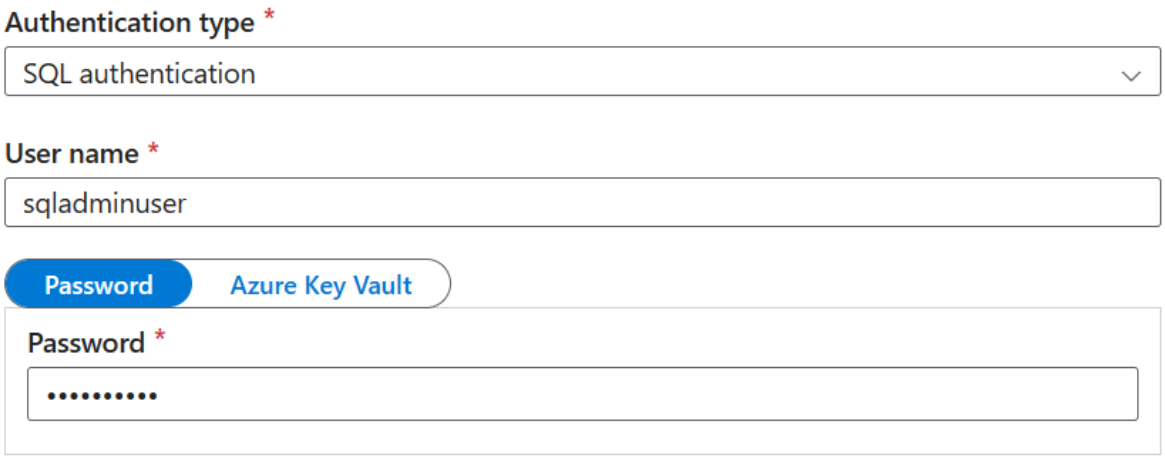
30. First you are going to give it the same name as your Synapse Workspace.



31. Now here you have to choose your subscription then your server name of Synapse workspace then choose your Pool db.

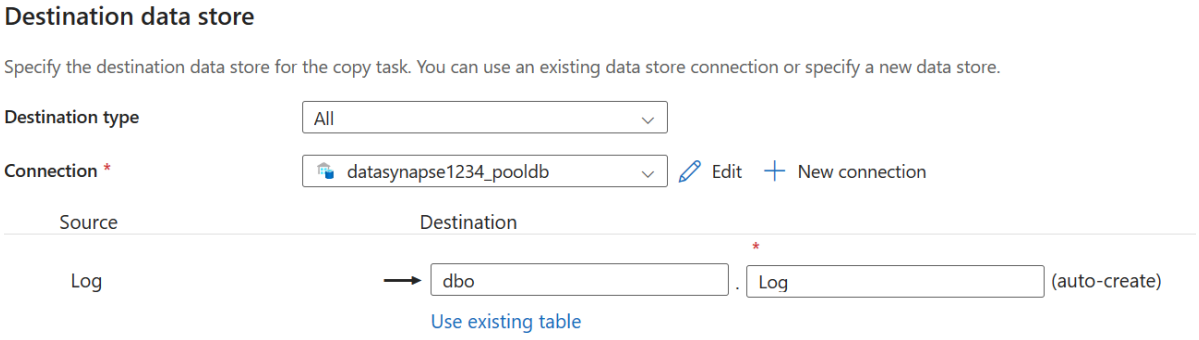


32. Then in the authentication type choose SQL Authentication and then give username of your Synapse Analytics and then give your password.

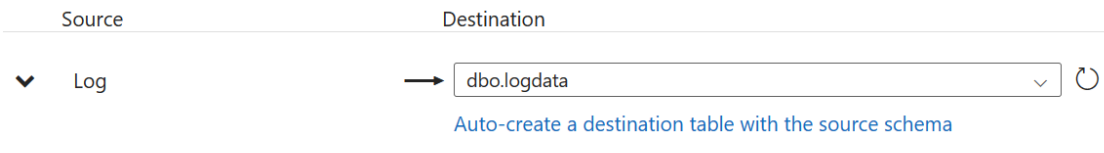


33. After that we will test our connection. Once your connection is successful then click on create.

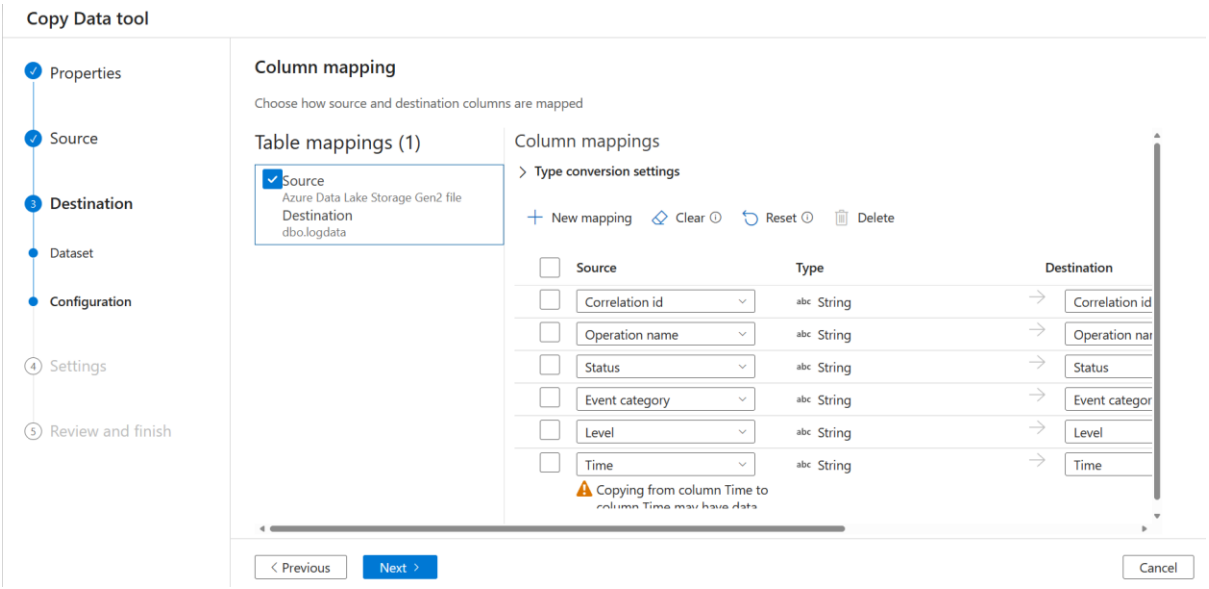
34. Now you have to choose your existing table for that click on use existing table.



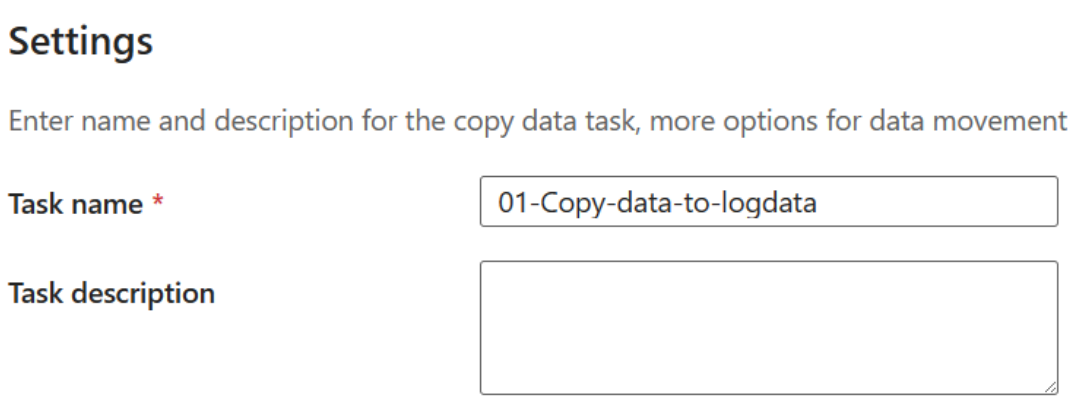
35. Below you can see that we have chosen our existing table. Then click on next.



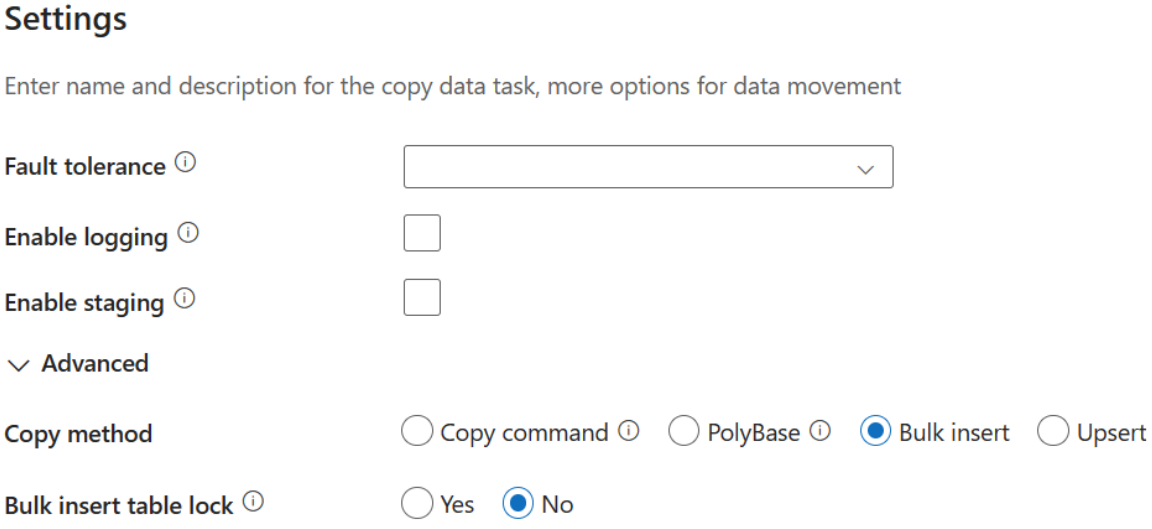
36. Now it will do a column mapping for your table. Click on next.



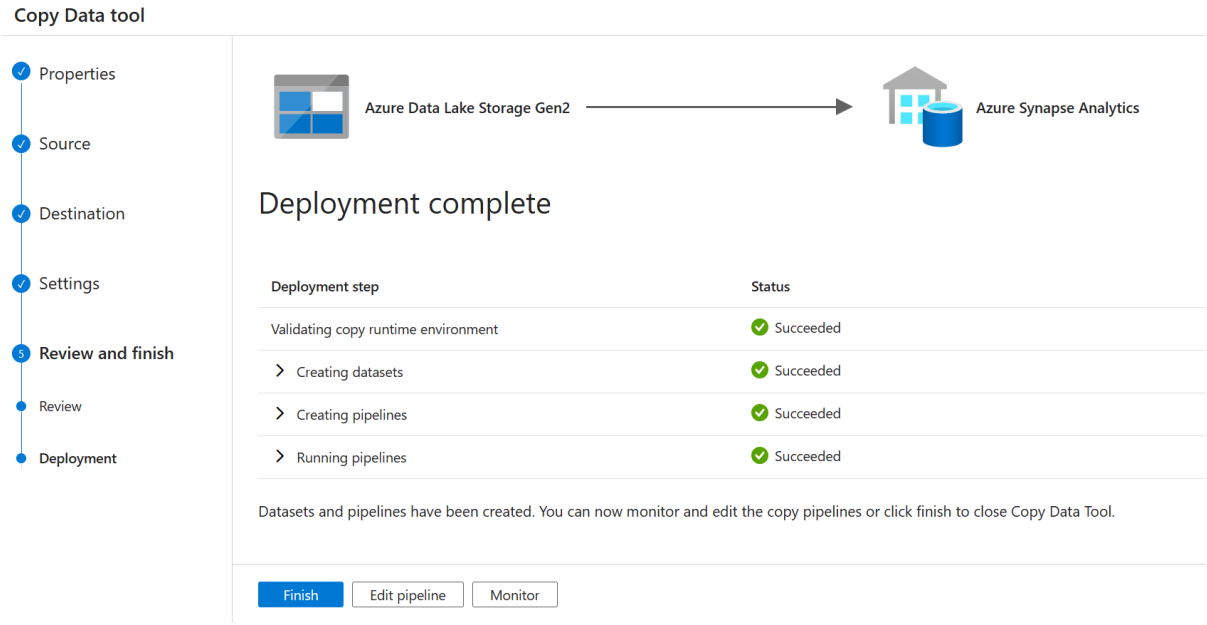
37. Then give it a name and scroll down.



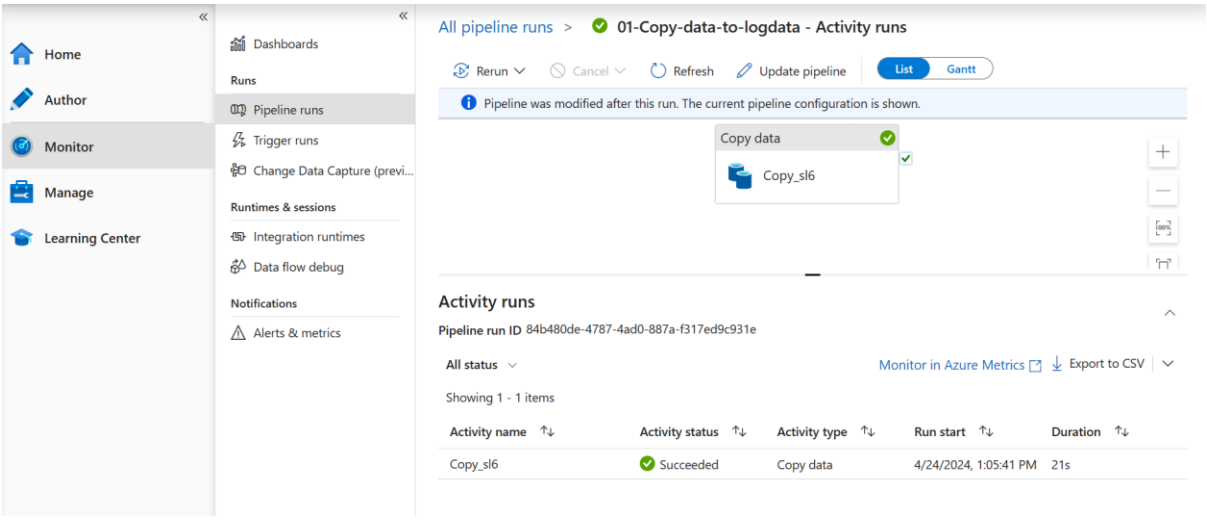
38. Here you have to disable staging and choose bulk insert.



39. After that click on next and move to the review page and deploy your pipeline. Once it is successfully deployed then click on finish.



40. Now move to the monitor section and here you can see that your pipeline run was successful.



41. Once it is done now move to SSMS there you have to run the Select query and check for your data.
42. Below you can see your data.

```
SELECT * FROM [logdata]
```

150 %

Results | Messages

	Correlation id	Operation name	Status	Event category	Level	Time
1	96600577-c906-4df2-8cd0-69c2872c7fdd	Create or Update Virtual Machine Extension	Started	Administrative	Informational	2023-04-17 06:31:00.723
2	c43307c3-015b-4c18-9556-456410c313de	Delete Network Security Group	Succeeded	Administrative	Informational	2023-04-17 06:07:51.210
3	70e3851e-079b-40d7-bb87-31f5386befcc	Delete Disk	Succeeded	Administrative	Informational	2023-04-17 06:04:33.950
4	cbbc3740-5009-4821-883c-cac8fd3c3659	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-17 05:46:03.813
5	cbbc3740-5009-4821-883c-cac8fd3c3659	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-17 05:46:02.207
6	f4c54d30-0b68-4c60-94fc-69349e63b788	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-17 05:44:07.787
7	304f186a-198f-4c52-8a1c-1baa98cec9cd	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-17 05:41:07.000
8	304f186a-198f-4c52-8a1c-1baa98cec9cd	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-17 05:41:05.563
9	bc53c993-3e88-4a7e-a4dd-30f01ca14790	Validate Deployment	Succeeded	Administrative	Informational	2023-04-17 05:33:32.583
10	a56c2662-433b-4cd6-ba1a-8f18c024271a	Delete Network Interface	Succeeded	Administrative	Informational	2023-04-16 14:13:41.823
11	77903c28-3455-4d4f-8e03-fbcdd13d6181	Create Flow Log	Succeeded	Administrative	Informational	2023-04-16 12:39:19.997
12	530772f7-90a4-4741-89ac-4941e02b2ca2	Create Connection Monitor	Started	Administrative	Informational	2023-04-16 11:47:31.090
13	a265ddd3-d275-46ec-9bc1-f2633cf37b90	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-16 06:18:40.863
14	420e3dc7-c154-49b1-9e49-f2f44de74cd6	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-16 06:14:42.030
15	420e3dc7-c154-49b1-9e49-f2f44de74cd6	Create or Update Virtual Machine Extension	Succeeded	Administrative	Informational	2023-04-16 06:07:46.280
16	7907a8a7-2b1c-4f22-b1f0-2c580435fde4	Health Event Updated	Updated	Resource Health	Informational	2023-04-16 06:04:43.103
17	cf800f9f-2782-4e7e-8060-3302d96fadf8	Delete Disk	Started	Administrative	Informational	2023-04-14 15:40:14.787
18	78e56a68-af47-42db-8d63-a0afa9930652	Get Network Interface Effective Security Groups	Started	Administrative	Informational	2023-04-14 15:23:11.450
19	a3ff6da1-818b-4449-b23b-c7b14a944513	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-14 15:07:56.157

Query executed successfully.

43. Now if you come back to the data factory and go to author you can see your pipeline which was created earlier and related to that there are several activities. Then in the datasets, you can see two datasets in place one is for the destination and the other is for the source.