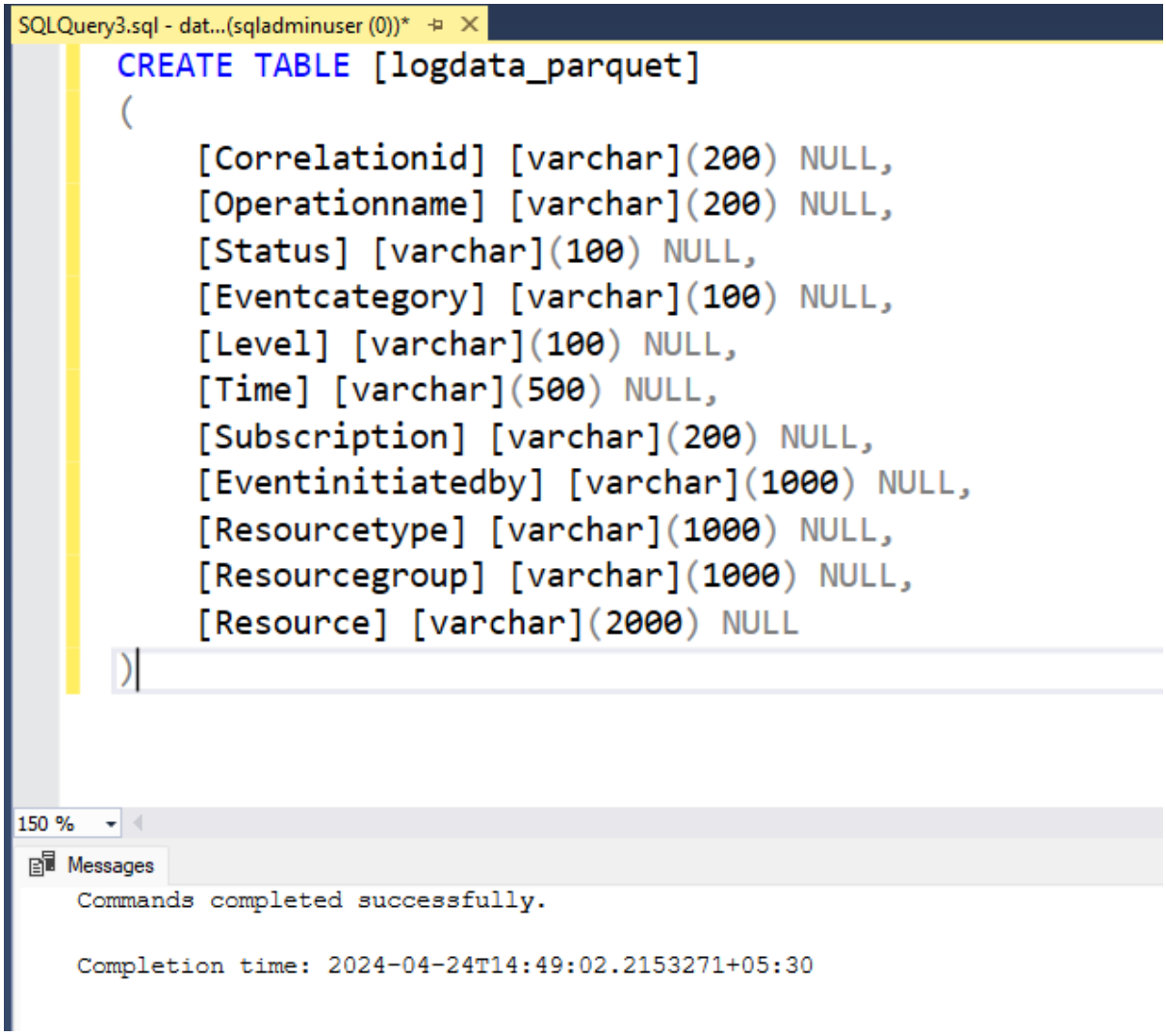
😊 Two-Step Process for the Pipeline

- In this lab, we're embarking on a two-step process within Azure Data Factory to facilitate efficient data transformation and ingestion. Our primary objective is to automate the conversion of data from a Parquet file format into a structured table within Azure Synapse Analytics, enabling seamless integration and analysis of data.
- To kickstart the process, we first create a table within the dedicated SQL pool of Azure Synapse Analytics using SQL Server Management Studio (SSMS). This step lays the foundation for our data pipeline, providing a target destination for the data extracted from the Parquet file.
- Moving forward, within the Azure Data Factory interface, we initiate the creation of a new pipeline. This pipeline serves as the framework for orchestrating the two-step data transformation process. Our initial step involves copying data from the Parquet file, residing in Azure Data Lake Storage, into our SQL table.
- Within the pipeline canvas, we configure a Copy Data activity to handle the data movement. We specify the Parquet file as the source dataset, employing a wildcard file path to accommodate potential variations in file names. This dynamic approach ensures flexibility in capturing all relevant data files generated in the Parquet container.
- For the destination, we create a new dataset representing the SQL table in the dedicated SQL pool. This dataset is linked to the target table within Azure Synapse Analytics, facilitating the seamless transfer of data from the Parquet file to the structured table. We opt for the bulk insert sink type to optimize data loading performance.
- To ensure data integrity and accuracy during the transformation process, we meticulously map the columns between the source Parquet file and the destination SQL table. This mapping exercise guarantees that data attributes are correctly aligned, preserving the integrity of the dataset.
- With the data movement and mapping configurations in place, we establish a dependency between the two activities within the pipeline. This dependency dictates that the second activity, responsible for copying data into the SQL table, only initiates upon successful completion of the initial data copying task. This sequential execution ensures a streamlined and error-resilient process flow.
- After validating and publishing the pipeline, we proceed to trigger its execution. The pipeline orchestrates the sequential execution of the two activities, seamlessly transferring data from the Parquet file to the SQL table within Azure Synapse Analytics.
- Upon successful execution, we verify the completion of the pipeline run and inspect the target SQL table to ensure the accurate ingestion of data. This final verification step validates the effectiveness of our data transformation pipeline, demonstrating its capability to automate complex data integration tasks within the Azure ecosystem.

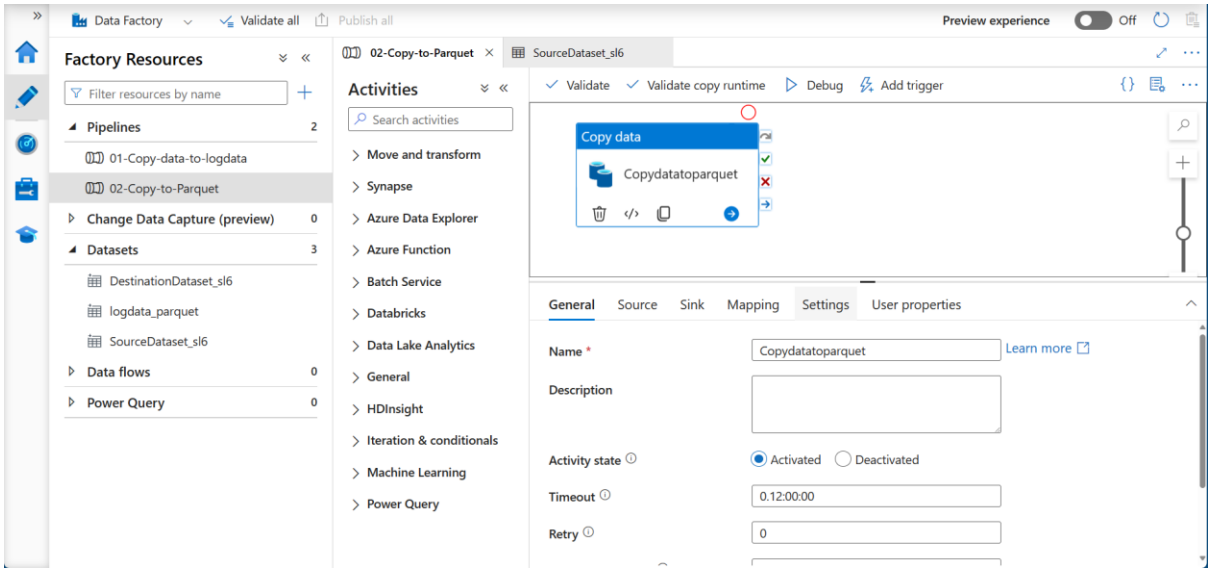
In summary, this lab empowers users to harness the capabilities of Azure Data Factory for orchestrating efficient data transformation and ingestion processes, ultimately facilitating data-driven decision-making and analytics within Azure Synapse Analytics.

😄 To begin with the Lab:

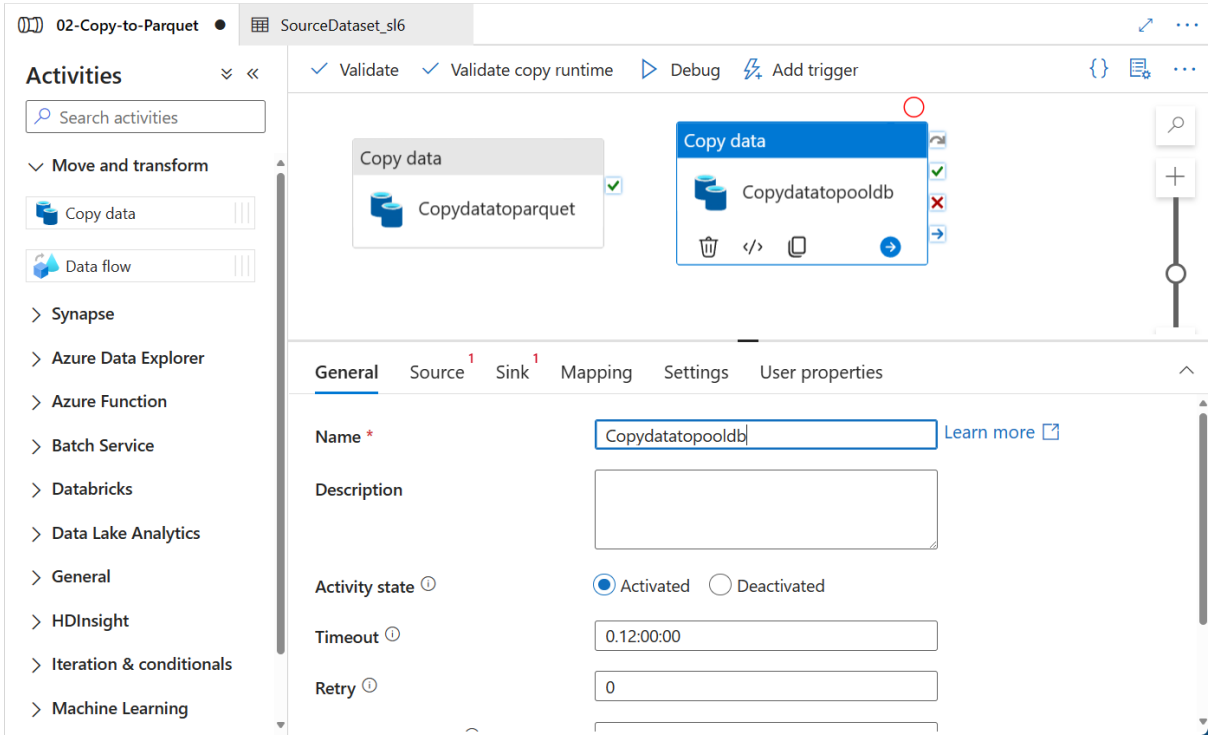
1. In the last lab we created a pipeline to copy data from a CSV-based file to a Parquet-based file.
2. Now in this lab we will use our parquet file to copy the data into the table.
3. First, we need to create a table for the parquet. Like we did in our first lab, we have to open SSMS and there we are going to create a table in our pool db.



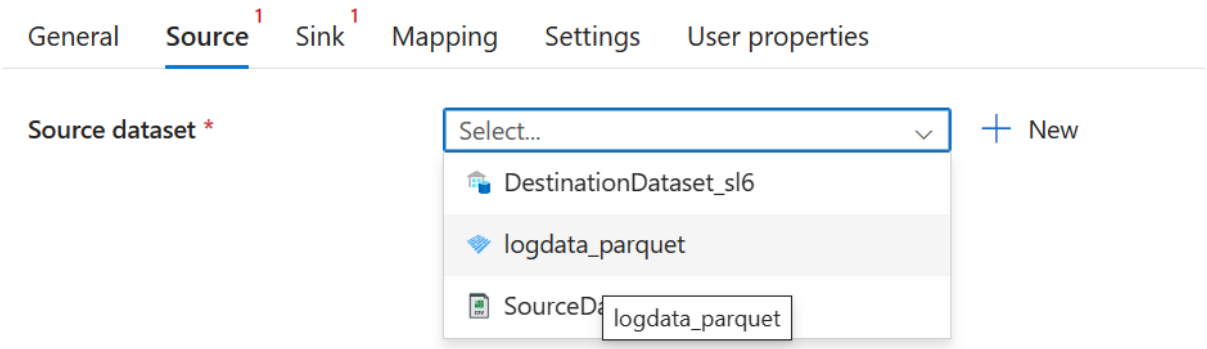
4. Now come back to our Data Factory wizard and there you can see that we have one activity in place which was to copy the data.



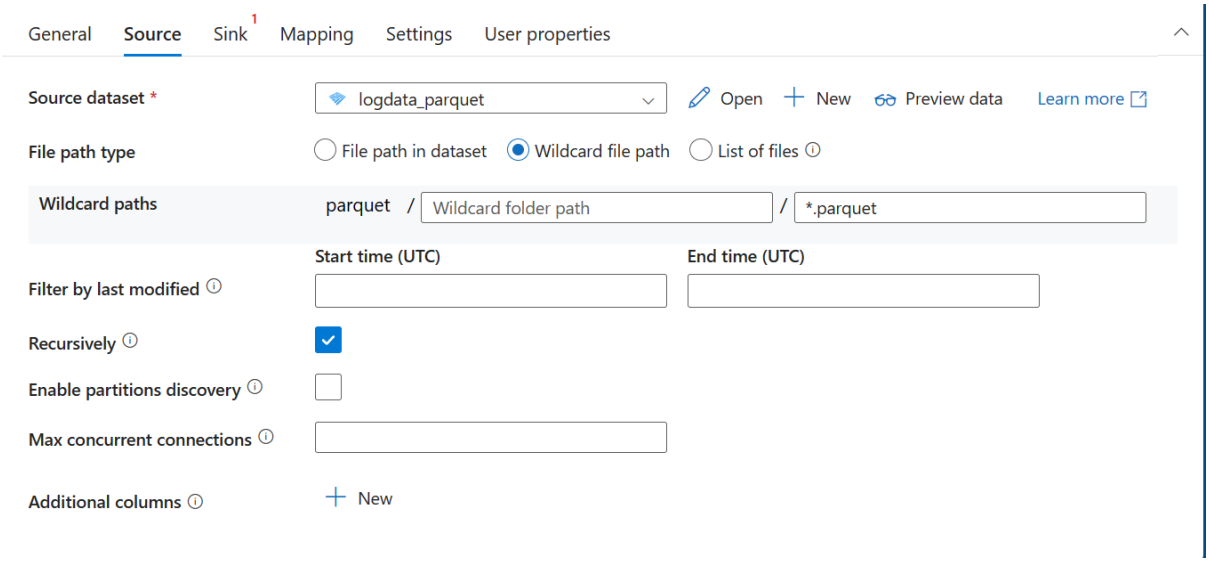
5. From the activities expand move and transform then drag the copy data tool to canvas.
6. Then give it a name.



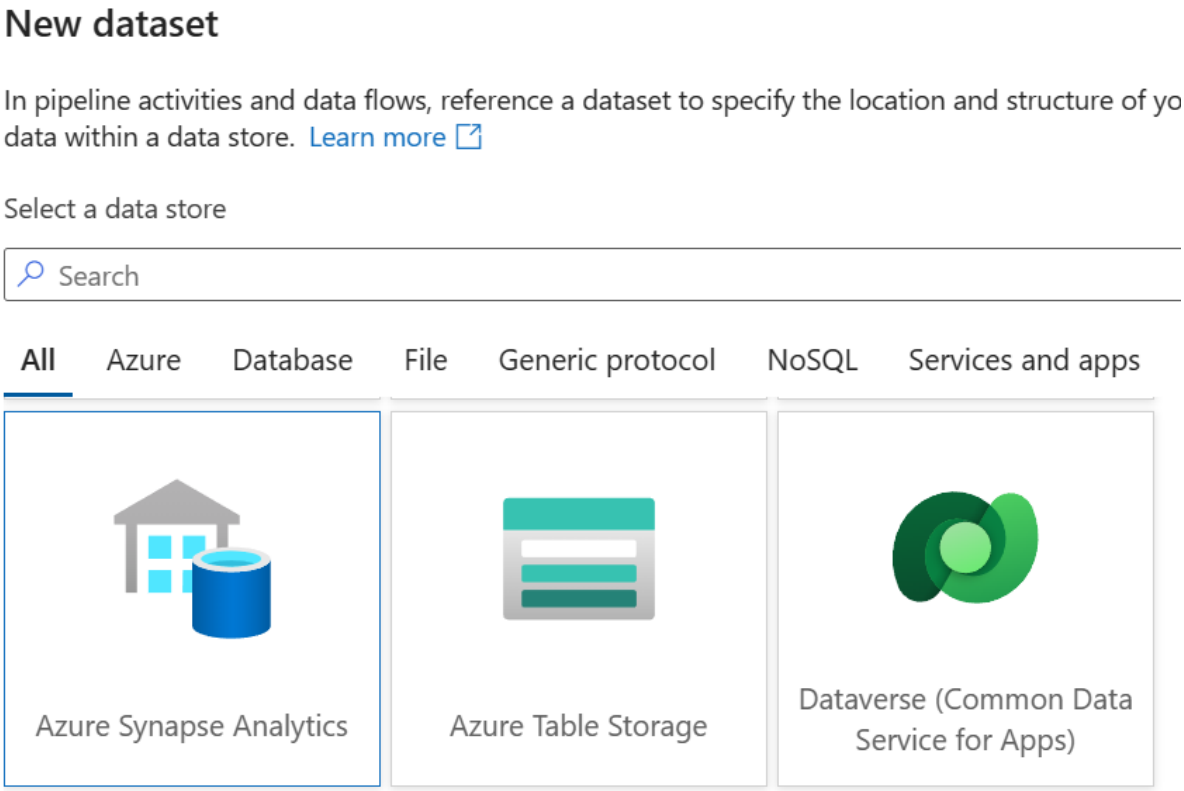
7. After that move to source and choose your logdata_parquet.



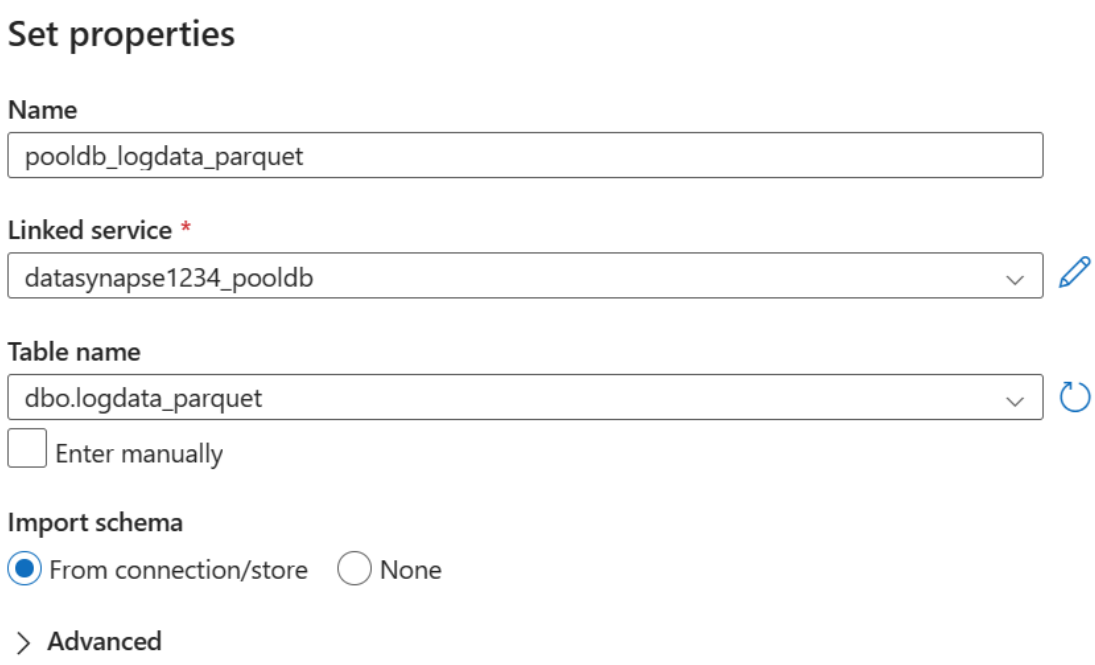
8. This time, we'll choose a wildcard file path because we want to take now whatever the generated Parquet-based files in the Parquet-based container to be copied onto my dedicated SQL pool.



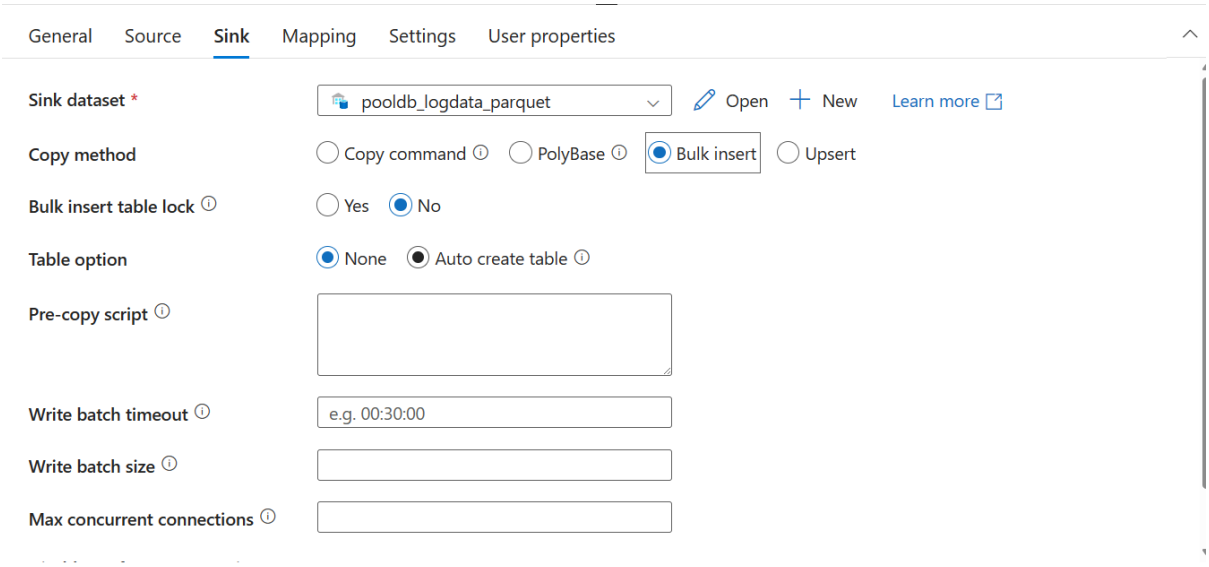
9. Then in the Sink we don’t have anything that points towards our pool db. So, we need to create a new synced dataset.
10. Now we need to choose Azure Synapse Analytics.



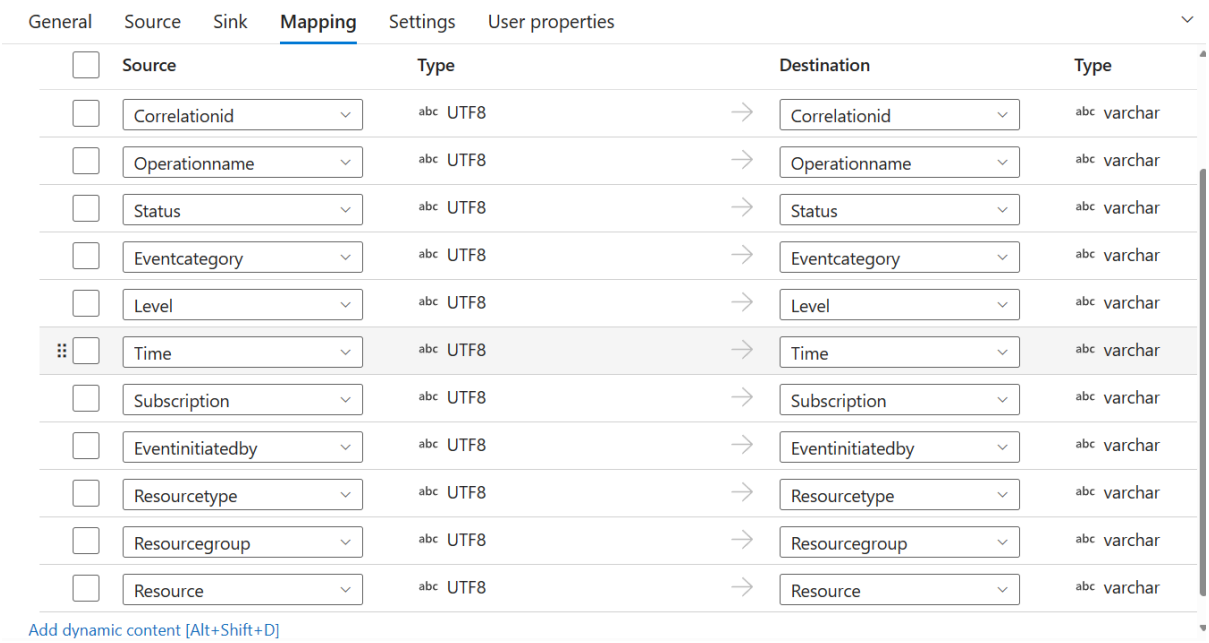
11. First, we will give it a name then we have to choose from a linked service which will be our pool db.
12. Then we have to choose our table name and then click on OK.



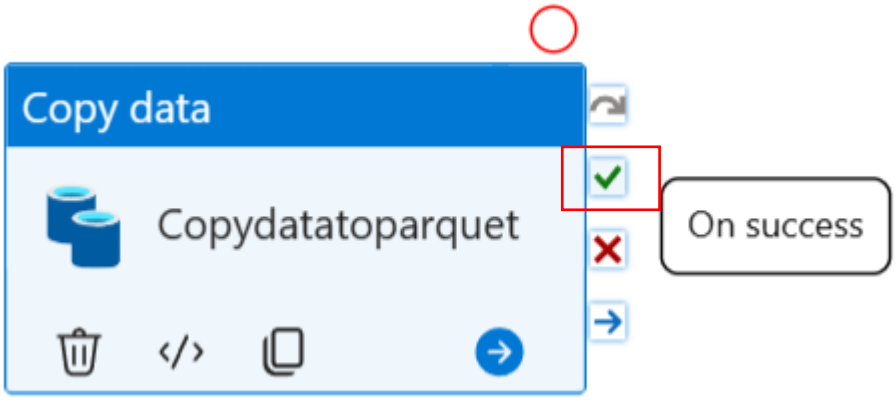
13. After that in the sink only you have to choose bulk insert.



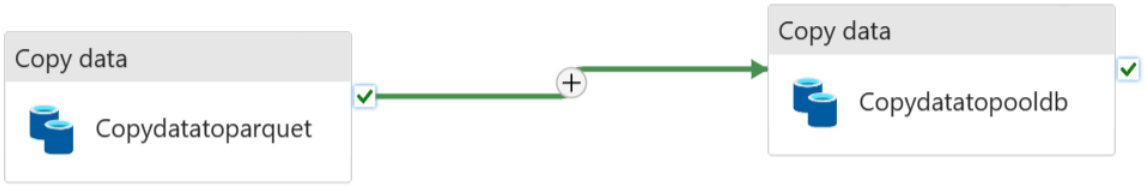
14. Then if you go to mappings and click on import them then you can see that this time we have the proper mapping.



15. Now we are going to connect both of our activities. Now in your Canvas if you take your mouse to this highlighted point shown below then you will get this on success. This means that when this activity is successful then only start the second activity.



16. Now we are going to connect them. Here you have to left click on your mouse from this checkpoint and drag it near your second activity. Then you will see this connection symbol between your primary activity and your secondary activity.

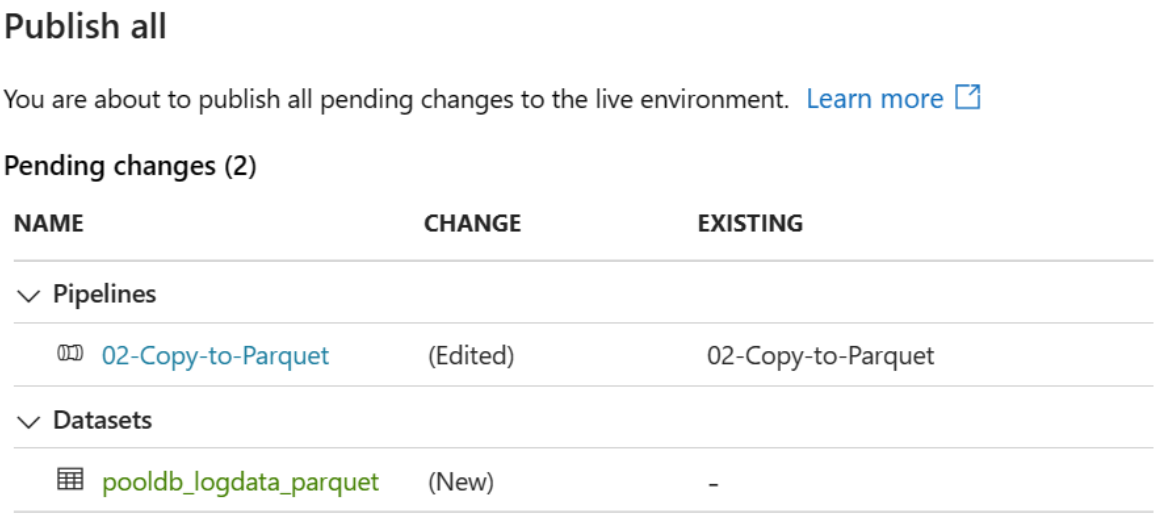


17. Please note here, that it's not that the output of this activity is going onto this activity. No, the output of this activity is going onto a Parquet-based container in Azure Data Lake. And then this activity will start, this activity will pick up the data from that Parquet-based container.
18. You're always saying that when this activity is completed successfully, it's a success, then go ahead and trigger the next activity.
19. Then you have to click on validate all. You can see that the validation has passed.

Your factory has been validated.

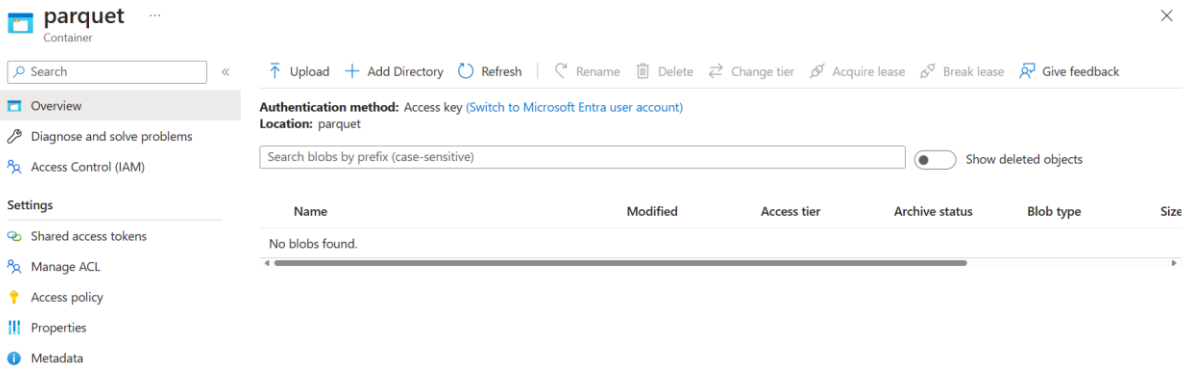
No errors were found.

20. Then click on publish all. It will publish all of your activities.

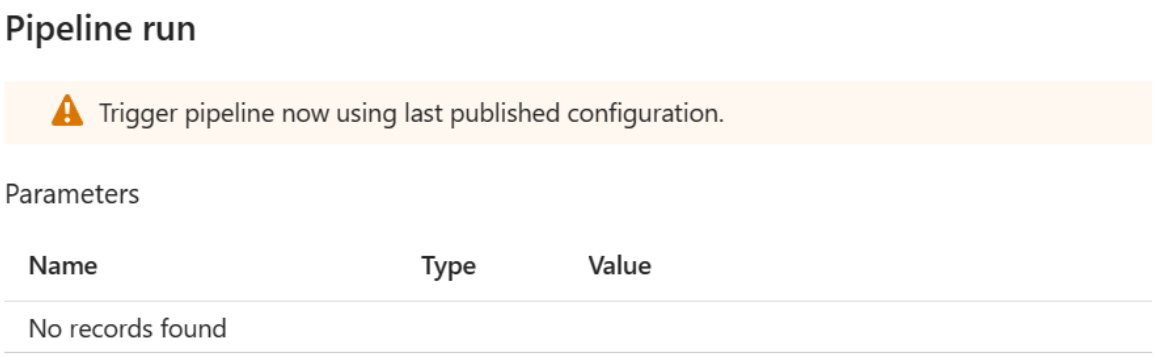


21. Below you can see that publishing has been completed.

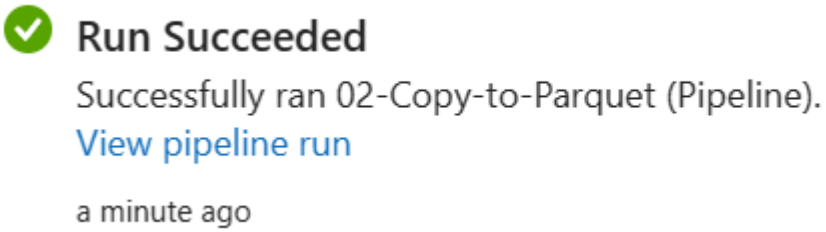
22. Now as we are doing both activities, it is better to delete our parquet file from our container.
23. So, go to your storage account and empty your parquet container.



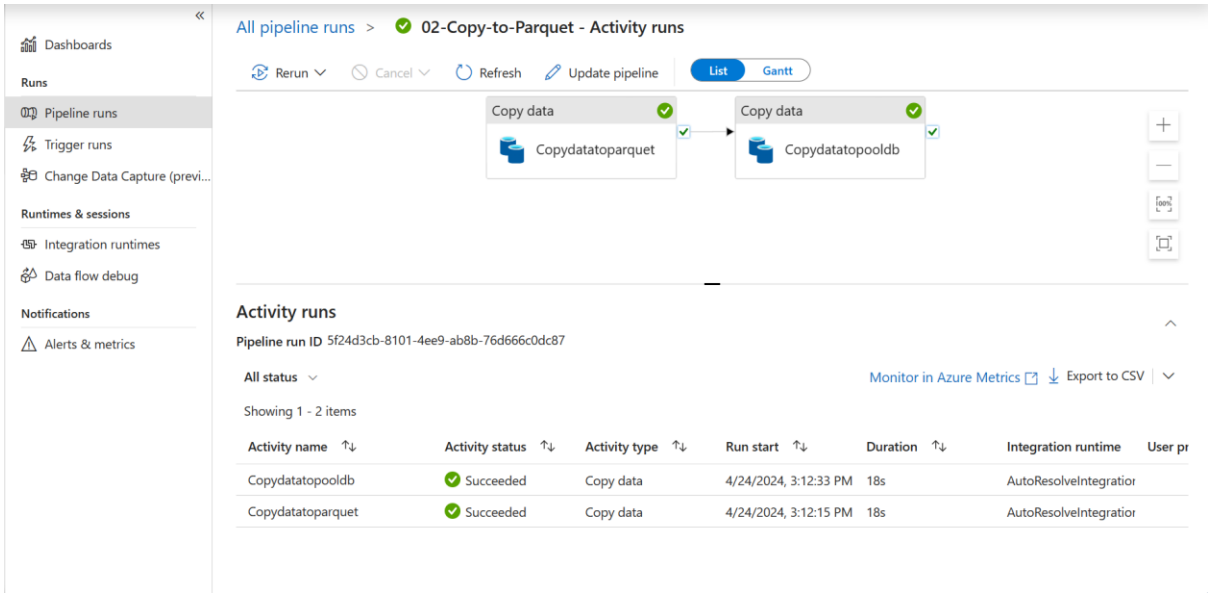
24. After that come back to Data Factory Wizard. Then click on Add trigger after that on the new trigger.
25. So now our pipeline consists of two activities. One is to first copy the log.csv data or convert it onto a log. parquet-based file and then copy those contents onto our pool db, our dedicated SQL pool.



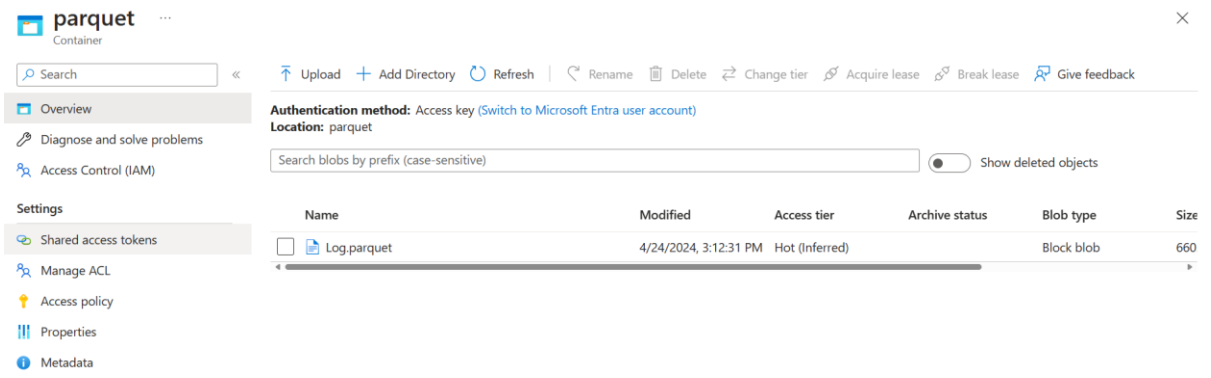
26. You can see that our pipeline run was successful. Now click on view pipeline run.



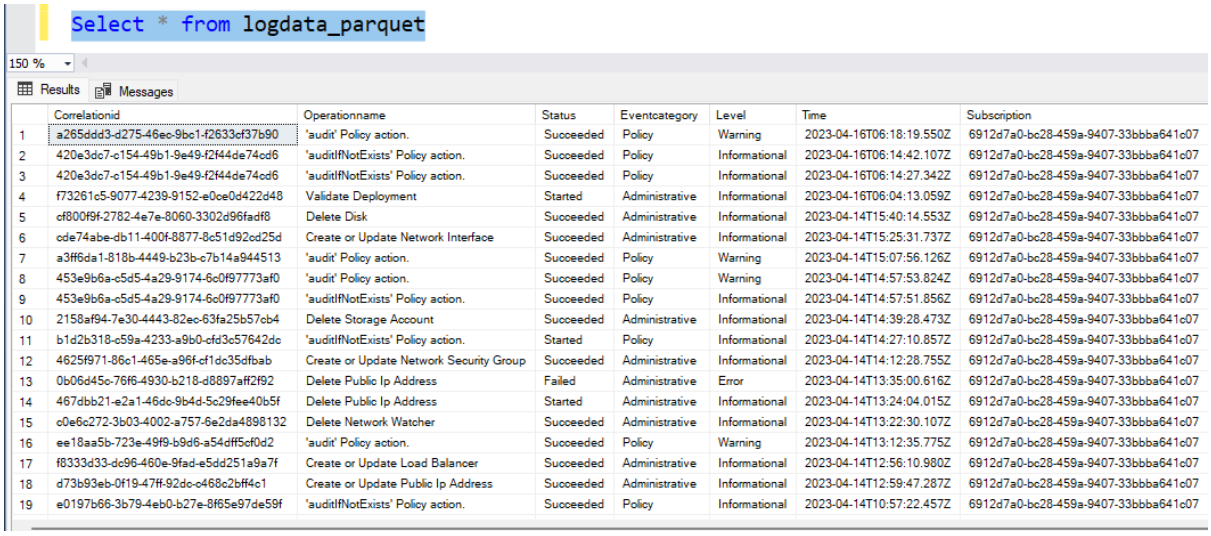
27. Here you can see both of them.



28. Now go back to your container. There you will see your parquet file.



29. Then go to SSMS and run the select command. And you can see your data accordingly.



```
Select * from logdata_parquet
```

150 %

Results | Messages

	Correlationid	Operationname	Status	Eventcategory	Level	Time	Subscription
1	a265ddd3-d275-46ec-9bc1-f2633cf37b90	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-16T06:18:19.550Z	6912d7a0-bc28-459a-9407-33bbba641c07
2	420e3dc7-c154-49b1-9e49-f2f44de74cd6	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-16T06:14:42.107Z	6912d7a0-bc28-459a-9407-33bbba641c07
3	420e3dc7-c154-49b1-9e49-f2f44de74cd6	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-16T06:14:27.342Z	6912d7a0-bc28-459a-9407-33bbba641c07
4	f73261c5-9077-4239-9152-e0ce0d422d48	Validate Deployment	Started	Administrative	Informational	2023-04-16T06:04:13.059Z	6912d7a0-bc28-459a-9407-33bbba641c07
5	cf800f9f-2782-4e7e-8060-3302d96fadf8	Delete Disk	Succeeded	Administrative	Informational	2023-04-14T15:40:14.553Z	6912d7a0-bc28-459a-9407-33bbba641c07
6	cde74abe-db11-400f-8877-8c51d92cd25d	Create or Update Network Interface	Succeeded	Administrative	Informational	2023-04-14T15:25:31.737Z	6912d7a0-bc28-459a-9407-33bbba641c07
7	a3ff6da1-818b-4449-b23b-c7b14a944513	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-14T15:07:56.126Z	6912d7a0-bc28-459a-9407-33bbba641c07
8	453e9b6a-c5d5-4a29-9174-6c0f97773af0	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-14T14:57:53.824Z	6912d7a0-bc28-459a-9407-33bbba641c07
9	453e9b6a-c5d5-4a29-9174-6c0f97773af0	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-14T14:57:51.856Z	6912d7a0-bc28-459a-9407-33bbba641c07
10	2158af94-7e30-4443-82ec-63fa25b57cb4	Delete Storage Account	Succeeded	Administrative	Informational	2023-04-14T14:39:28.473Z	6912d7a0-bc28-459a-9407-33bbba641c07
11	b1d2b318-c59a-4233-a9b0-cfd3c57642dc	'auditIfNotExists' Policy action.	Started	Policy	Informational	2023-04-14T14:27:10.857Z	6912d7a0-bc28-459a-9407-33bbba641c07
12	4625f971-86c1-465e-a96f-cf1dc35dfbab	Create or Update Network Security Group	Succeeded	Administrative	Informational	2023-04-14T14:12:28.755Z	6912d7a0-bc28-459a-9407-33bbba641c07
13	0b06d45c-76f6-4930-b218-d8897aff2f92	Delete Public Ip Address	Failed	Administrative	Error	2023-04-14T13:35:00.616Z	6912d7a0-bc28-459a-9407-33bbba641c07
14	467dbb21-e2a1-46dc-9b4d-5c29fee40b5f	Delete Public Ip Address	Started	Administrative	Informational	2023-04-14T13:24:04.015Z	6912d7a0-bc28-459a-9407-33bbba641c07
15	c0e6c272-3b03-4002-a757-6e2da4898132	Delete Network Watcher	Succeeded	Administrative	Informational	2023-04-14T13:22:30.107Z	6912d7a0-bc28-459a-9407-33bbba641c07
16	ee18aa5b-723e-49f9-b9d6-a54dff5cf0d2	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-14T13:12:35.775Z	6912d7a0-bc28-459a-9407-33bbba641c07
17	f8333d33-dc96-460e-9fad-e5dd251a9a7f	Create or Update Load Balancer	Succeeded	Administrative	Informational	2023-04-14T12:56:10.980Z	6912d7a0-bc28-459a-9407-33bbba641c07
18	d73b93eb-0f19-47ff-92dc-c468c2bff4c1	Create or Update Public Ip Address	Succeeded	Administrative	Informational	2023-04-14T12:59:47.287Z	6912d7a0-bc28-459a-9407-33bbba641c07
19	e0197b66-3b79-4eb0-b27e-8f65e97de59f	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-14T10:57:22.457Z	6912d7a0-bc28-459a-9407-33bbba641c07