

Filtering Rows

1. In this lab we are going to filter some of our rows from the Parquet table.
2. As if you will look at the data that we have in Parquet file we can see that in the Resource group column some of the data is Null.
3. So, we used the below command and we get 191 rows that are null

```
SELECT * FROM [logdata_parquet]
WHERE Resourcegroup IS NULL
```

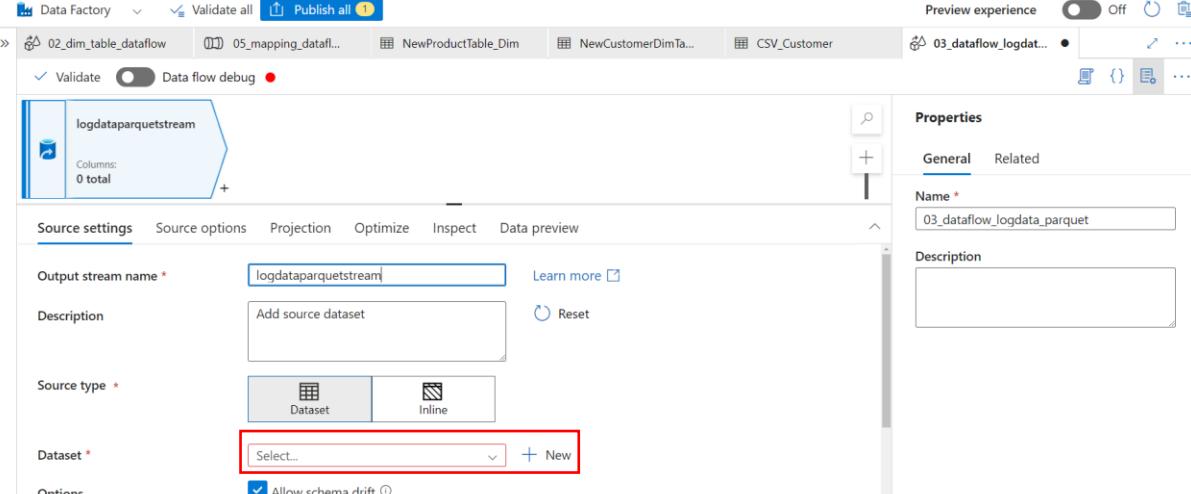
Resourcetype	Resourcegroup	Resource
NULL	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07
NULL	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
NULL	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
Microsoft.RecoveryServices/locations	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
NULL	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
NULL	NULL	/SUBSCRIPTIONS/6912D7A0-BC28-459A-9407-33BBBA641...
Microsoft.RecoveryServices/locations	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
NULL	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
NULL	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
Microsoft.RecoveryServices/locations	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
Microsoft.Network/locations	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
Microsoft.RecoveryServices/locations	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
NULL	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...
Microsoft.RecoveryServices/locations	NULL	/subscriptions/6912d7a0-bc28-459a-9407-33bbba641c07/pro...

| datasynapse1234.sql.azurestorage | sqladminuser (0) | demodb | 00:00:01 | 191 rows

4. As you can see below this is the data which we'll get when we remove the Null values.

-- Original Table - 19,229 rows
 -- With NULL values - 191
 -- Rows in the end - 19,038 rows

5. Now move to Data Factory wizard and create a new data flow.
6. Here first we will give name to our data flow then we choose to create a new resource then name it.
7. After that you need to choose the dataset.



The screenshot shows the Azure Data Factory interface for creating a new dataset. The top navigation bar includes 'Data Factory', 'Validate all', 'Publish all', and various validation status indicators. The main workspace shows a dataset named 'logdataparquetstream' with one column. The 'Properties' panel on the right shows the dataset's name as '03_dataflow_logdata_parquet'. The 'Source settings' tab is selected, showing fields for 'Output stream name' (set to 'logdataparquetstream'), 'Description' (set to 'Add source dataset'), 'Source type' (set to 'Dataset'), and a dropdown for 'Dataset' which is highlighted with a red box. Other tabs include 'Source options', 'Projection', 'Optimize', 'Inspect', and 'Data preview'. The 'Dataset' dropdown menu is open, showing 'Select...' and '+ New' options. A checkbox for 'Allow schema drift' is checked.

8. Now we need to create the dataset for that click on new.



9. Then you have to choose your Azure data lake gen2 storage account, then you need to choose Parquet as your file format and after that in the set properties option you need to give it a name choose your linked service and give the file path for your parquet container and the file stored in it.

Set properties

Name

NewParquetdataset

Linked service *

sqlstorage1010_service



File path

parquet

/ Directory

/ Log.parquet



Import schema

From connection/store From sample file None

> Advanced

10. After that go back to SSMS and delete the data in your Parquet file.

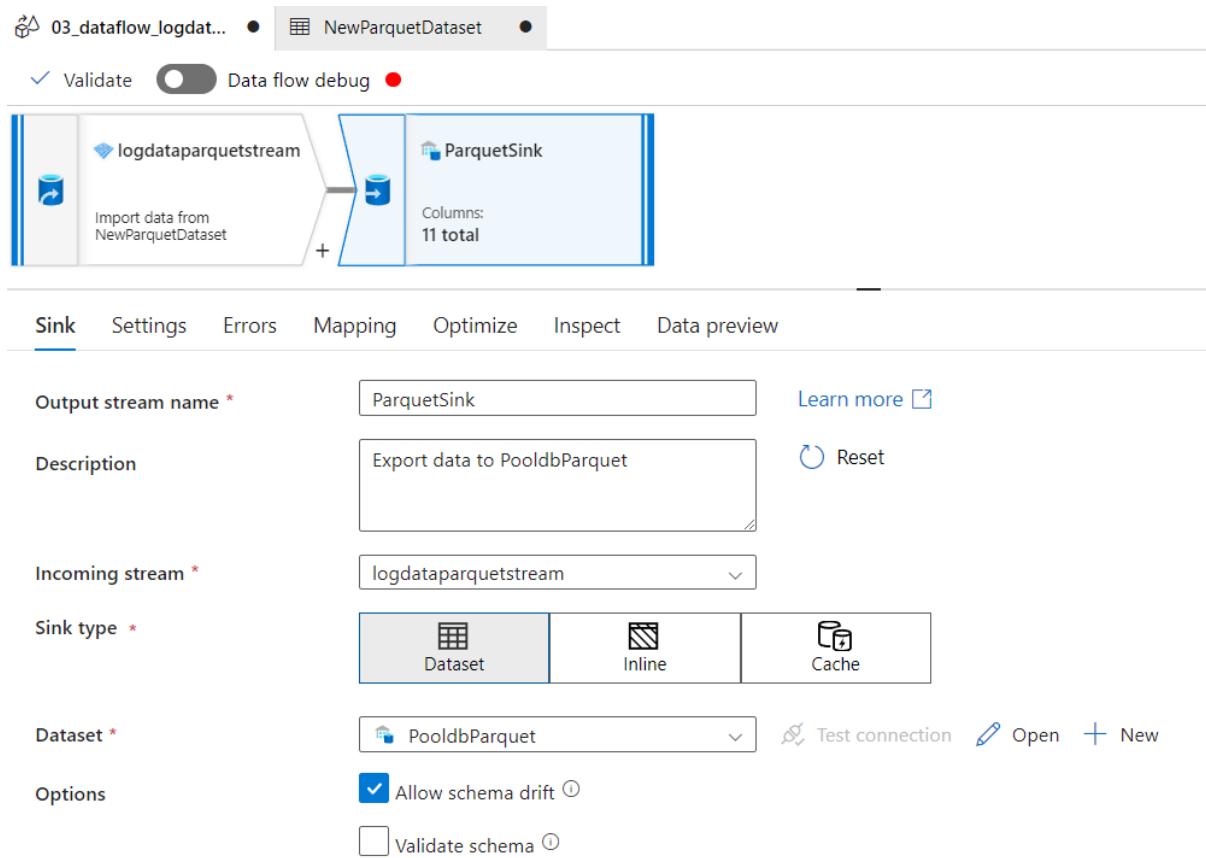
The screenshot shows a SQL Server Management Studio (SSMS) window titled "SQLQuery1.sql - dat... (sqladminuser (0))". Inside the query editor, the following T-SQL command is written:

```
DELETE FROM [logdata_parquet]
```

Below the query, the status bar displays the completion time: "Completion time: 2024-05-02T19:39:56.6162423+05:30".

11. Then you need to create a sink in the data flow.

12. Now in your sink give it a name then for the dataset click on new and then choose Synapse Analytics.



13. After that in the set properties you need to give it a name then in the linked service choose your pool db. Then you have to choose your parquet table.

Set properties

Name

PooldbParquet

Linked service *

synapse_pooldb

Select from existing table New table

Table name

Select...

Enter manually

Import schema

From connection/store None

> Advanced

14. Then click on the highlighted plus icon in your dataflow and choose filter option from the drop down menu.



15. Then give your filter a name and in the filter on option click on open expression builder.

Filter settings Optimize Inspect Data preview

Output stream name * [Learn more](#)

Description [Reset](#)

Incoming stream *

Filter on * ANY
[Open expression builder](#)

16. In the expression builder you have to write the same expression as shown below then click on save and finish.

Dataflow expression builder

FilterNullValues

Expression

```
!(isNull({Resourcegroup}))
```

17. Then just click on publish all and then go to the pipelines and create a new pipeline.

Publish all

You are about to publish all pending changes to the live environment. [Learn more](#)

Pending changes (3)

NAME	CHANGE	EXISTING
▽ Datasets		
NewParquetDataset	(New)	-
PooldbParquet	(New)	-
▽ Data flows		
03_dataflow_logdata_parq...	(New)	-

18. Create a pipeline for dataflow and then give it a name then choose your dataflow.

The screenshot shows the Azure Data Factory pipeline editor. On the left, there's a sidebar with various service options like Data Factory, Copy data, Data flow, Synapse, Azure Data Explorer, etc. The main area shows a pipeline named '03_dataflow_logdata_parquet'. This pipeline contains a single 'Data flow' activity, which is currently selected. The 'Settings' tab for the data flow is open, showing the following configuration:

- Data flow: 03_dataflow_logdata_parquet
- Run on (Azure IR): AutoResolveIntegrationRuntime
- Compute size: Small
- Logging level: Verbose

The 'General' tab for the pipeline shows the name is 'FilterParquetPipeline'. There are tabs for 'Properties', 'General', and 'Related' on the right side.

19. In staging choose your linked service and the container.

This screenshot shows the configuration for a 'Staging linked service'. The 'Staging linked service' dropdown is set to 'sqlstorage1010_service'. Below it, the 'Staging storage folder' is set to 'staging / Directory'. There are 'Edit' and 'New' buttons for managing linked services. A 'Browse' button is also present next to the folder path.

20. After that just publish your pipeline and click on trigger now.

21. Below you can see that our pipeline run was successful.

All pipeline runs > ✓ FilterParquetPipeline - Activity runs

Rerun Cancel Refresh Update pipeline List Gantt

Data flow DataflowFilter

Activity runs

Pipeline run ID: ce10e531-ba2b-4ce0-a3af-fb07c3c92bd5

All status ▾ Monitor in Azure Metrics Export to CSV

Showing 1 - 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
DataflowFilter	✓ Succeeded	Data flow	5/2/2024, 7:58:40 PM	3m 35s	AutoResolveIntegration		79161173-36dc-4158-96e

22. Now if we go to our SSMS and run the select statement for the parquet file then we'll get the filtered data.

23. As expected, we get the data and the number of rows are 19038

SQLQuery1.sql - dat... (sqladminuser (0))*

```
Select * From [logdata_parquet]
```

Results Messages

Correlationid	Operationname	Status	Eventcategory	Level	Time
1 76997d4c-0eb4-44f0-8f6b-d75af38942b5	Create or Update Load Balancer	Accepted	Administrative	Informational	2023-04-14T13:11:21.348Z
2 ee18aa5b-723e-49f9-b9d6-a54dff5cf0d2	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-14T13:12:35.463Z
3 ee18aa5b-723e-49f9-b9d6-a54dff5cf0d2	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-14T13:12:35.697Z
4 ee18aa5b-723e-49f9-b9d6-a54dff5cf0d2	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-14T13:12:35.697Z
5 ee18aa5b-723e-49f9-b9d6-a54dff5cf0d2	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-14T13:12:35.697Z
6 ee18aa5b-723e-49f9-b9d6-a54dff5cf0d2	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-14T13:12:35.713Z
7 ee18aa5b-723e-49f9-b9d6-a54dff5cf0d2	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-14T13:12:35.760Z
8 ee18aa5b-723e-49f9-b9d6-a54dff5cf0d2	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-14T13:12:35.760Z
9 ee18aa5b-723e-49f9-b9d6-a54dff5cf0d2	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-14T13:12:35.838Z
10 e57cb6fa-0e18-495f-a3b4-6db550bb13d4	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-06T11:53:17.108Z
11 e57cb6fa-0e18-495f-a3b4-6db550bb13d4	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-06T11:53:17.108Z
12 e57cb6fa-0e18-495f-a3b4-6db550bb13d4	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-06T11:53:17.108Z
13 e57cb6fa-0e18-495f-a3b4-6db550bb13d4	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-06T11:53:17.108Z
14 e57cb6fa-0e18-495f-a3b4-6db550bb13d4	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-06T11:53:17.124Z
15 e57cb6fa-0e18-495f-a3b4-6db550bb13d4	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-06T11:53:17.124Z
16 e57cb6fa-0e18-495f-a3b4-6db550bb13d4	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-06T11:53:17.139Z
17 e57cb6fa-0e18-495f-a3b4-6db550bb13d4	'auditIfNotExists' Policy action.	Succeeded	Policy	Informational	2023-04-06T11:53:17.233Z
18 e160-110-1120-1112-171-11-2-170	Process View Generated	Accepted	Administrative	Informational	2023-04-06T11:53:17.233Z

Query executed successfully. | datasynapse1234.sql.azuresy... | sqladminuser (0) | demodb | 00:00:00 | 19,038 rows