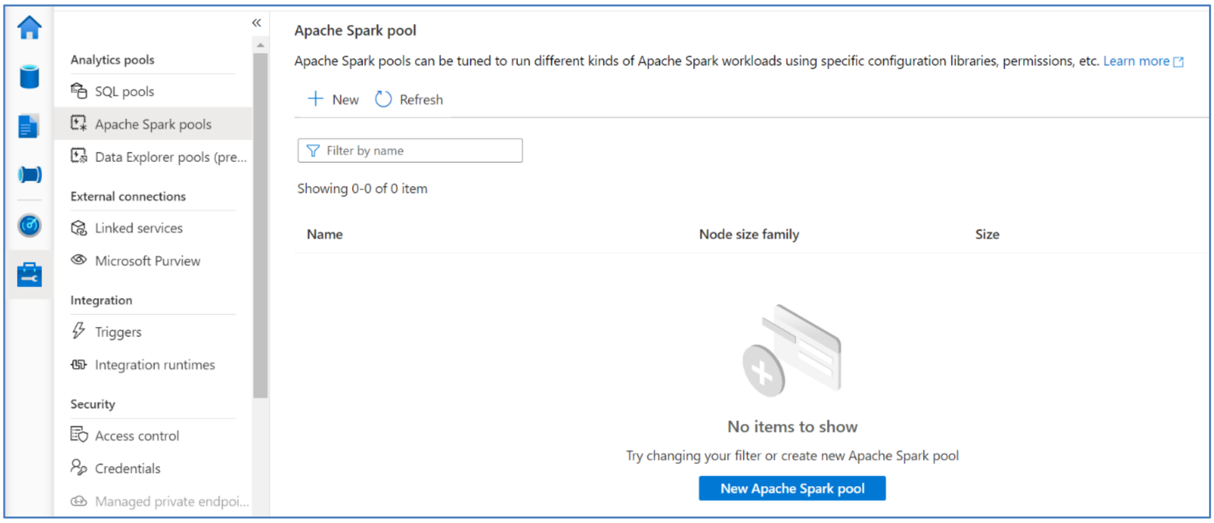
SPARK POOL

Step 1

Create Spark pool

Manage → Apache Spark Pool → New



Apache Spark pool

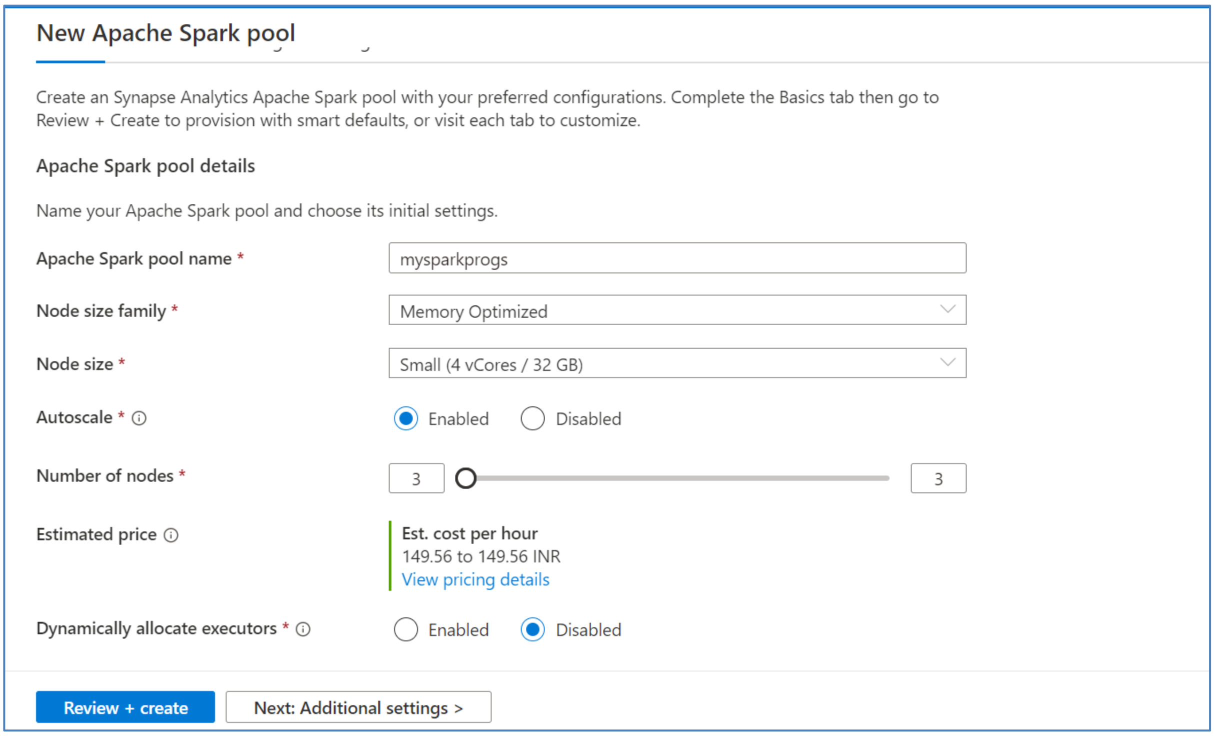
Apache Spark pools can be tuned to run different kinds of Apache Spark workloads using specific configuration libraries, permissions, etc. Learn more

Showing 0-0 of 0 item

Name	Node size family	Size

No items to show

Try changing your filter or create new Apache Spark pool



New Apache Spark pool

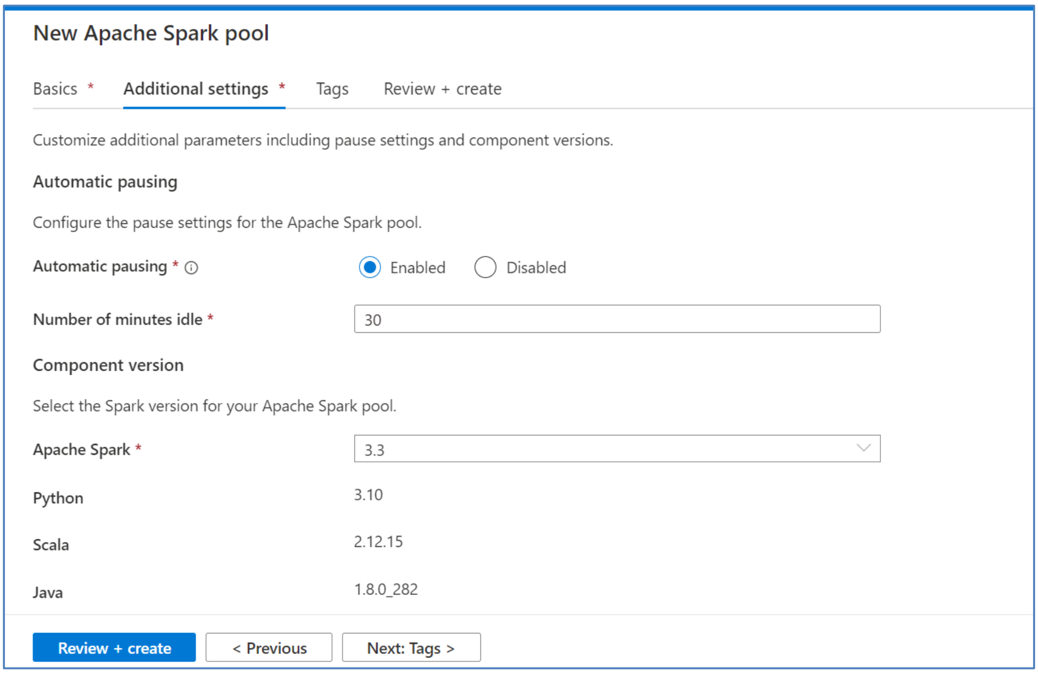
Create an Synapse Analytics Apache Spark pool with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize.

Apache Spark pool details

Name your Apache Spark pool and choose its initial settings.

Apache Spark pool name *	mysparkprogs
Node size family *	Memory Optimized
Node size *	Small (4 vCores / 32 GB)
Autoscale * ⓘ	● Enabled ○ Disabled
Number of nodes *	3 — 3
Estimated price ⓘ	Est. cost per hour 149.56 to 149.56 INR View pricing details
Dynamically allocate executors * ⓘ	○ Enabled ● Disabled

Under Advanced settings change the idle time to 30 mins.



New Apache Spark pool

Basics * | **Additional settings** * | Tags | Review + create

Customize additional parameters including pause settings and component versions.

Automatic pausing

Configure the pause settings for the Apache Spark pool.

Automatic pausing * ⓘ ● Enabled ○ Disabled

Number of minutes idle * 30

Component version

Select the Spark version for your Apache Spark pool.

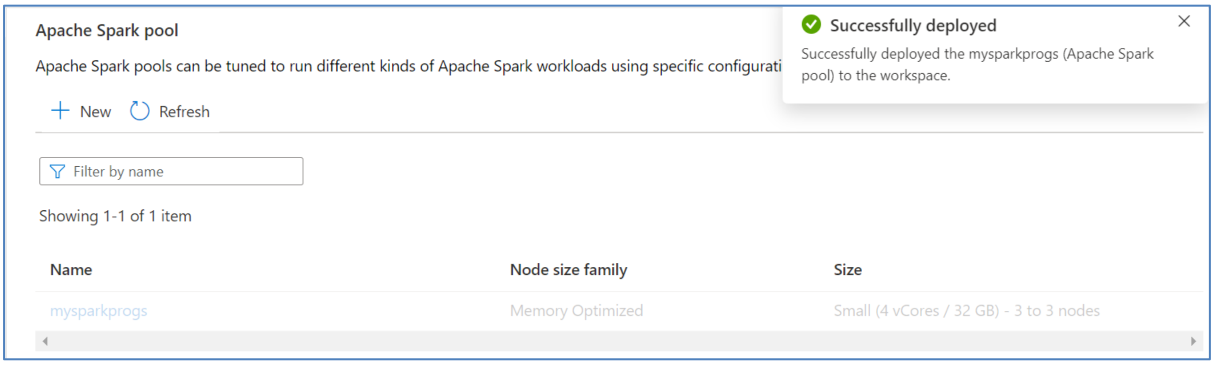
Apache Spark * 3.3

Python 3.10

Scala 2.12.15

Java 1.8.0_282

Click on **create**



Apache Spark pool

Apache Spark pools can be tuned to run different kinds of Apache Spark workloads using specific configurati

Successfully deployed

Successfully deployed the mysparkprogs (Apache Spark pool) to the workspace.

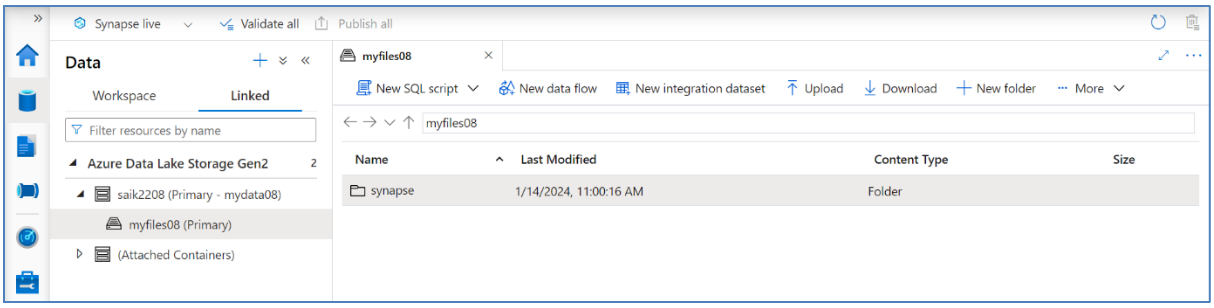
Showing 1-1 of 1 item

Name	Node size family	Size
mysparkprogs	Memory Optimized	Small (4 vCores / 32 GB) - 3 to 3 nodes

Step 2

Create folder by name "**mydata**" under the container and upload the **basestations_parquet** file.

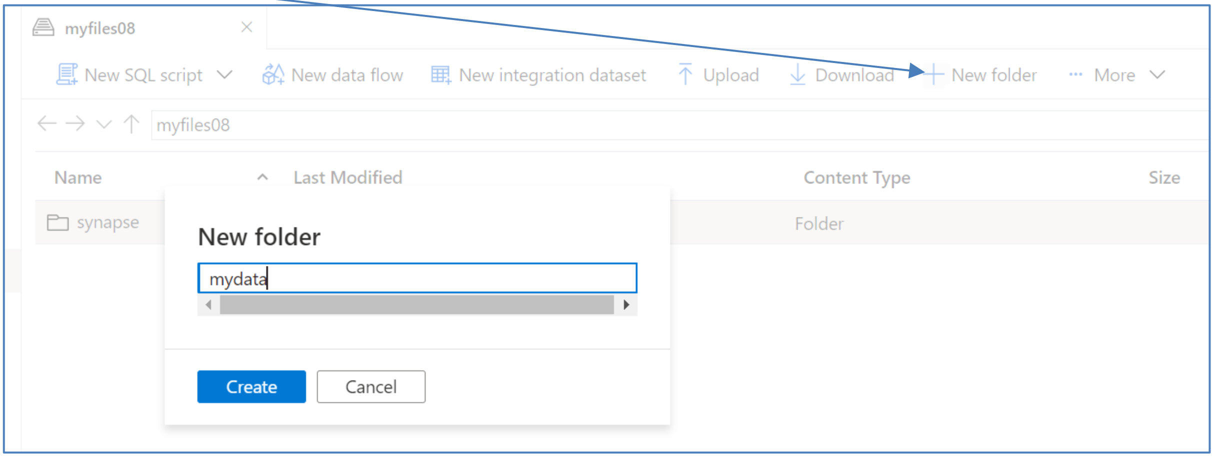
Go to **Data Hub –> Linked → Expand the Azure data lake storage →**



Name	Last Modified	Content Type	Size
synapse	1/14/2024, 11:00:16 AM	Folder	

Select the container

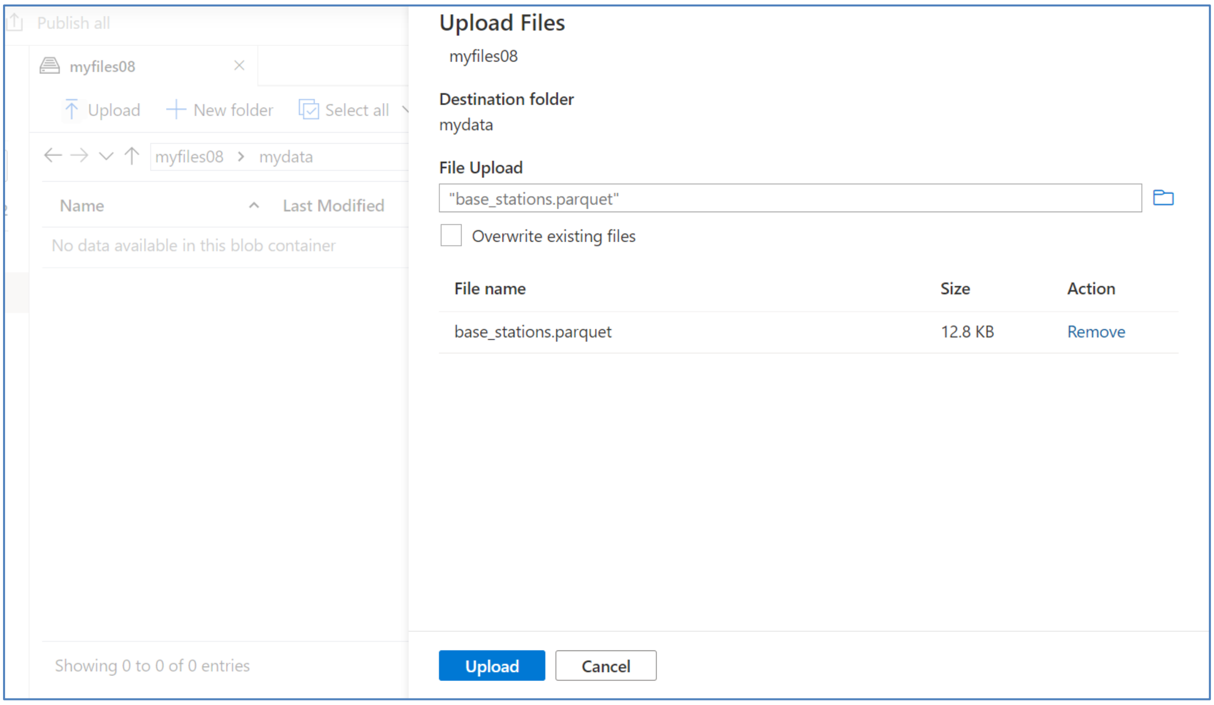
Click on **New folder** from top menu and create the folder.



New folder

mydata

Now double click on the folder and click on the **upload** and select the **base_stations.parquet** file to upload.



Upload Files

myfiles08

Destination folder

mydata

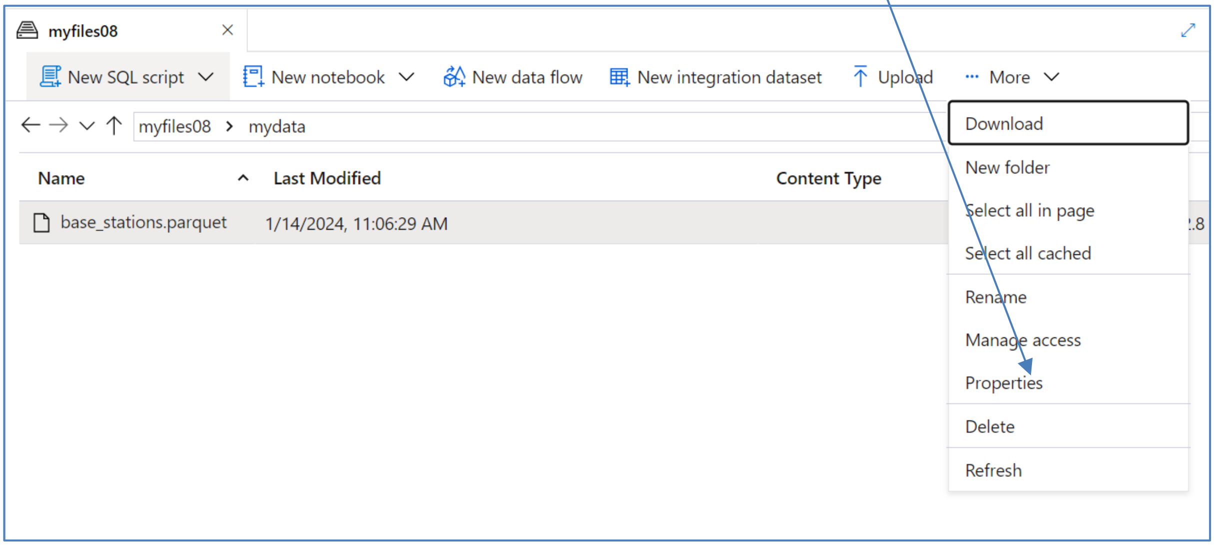
File Upload

"base_stations.parquet"

□ Overwrite existing files

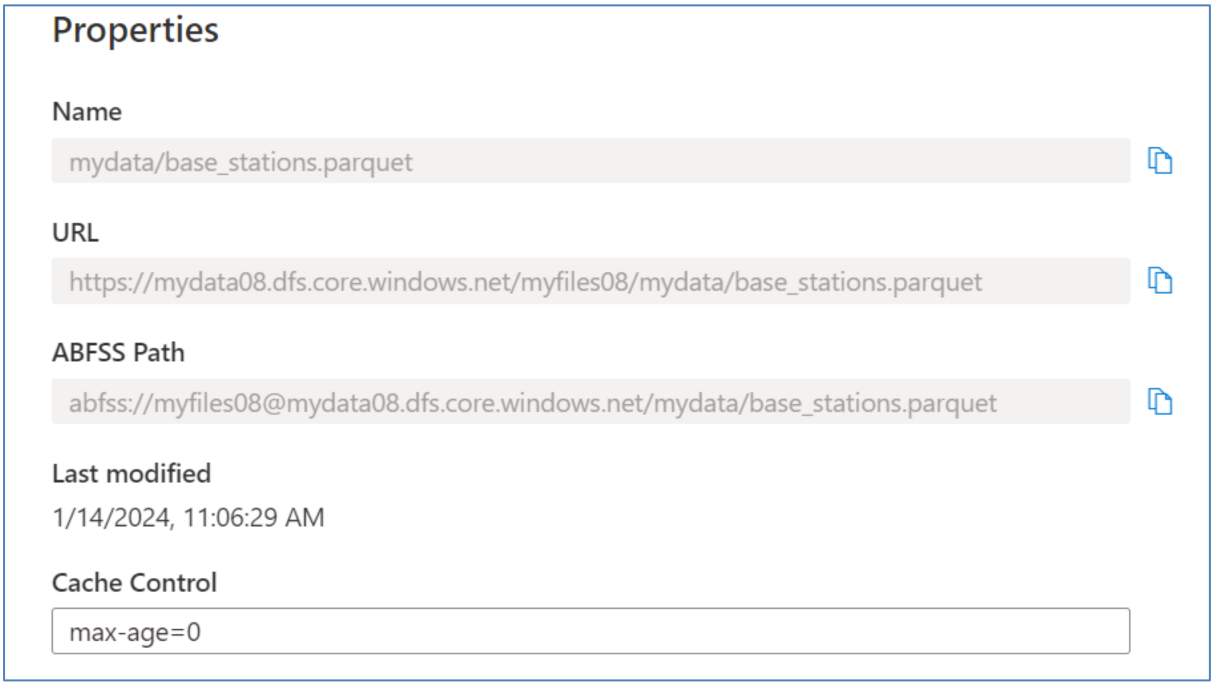
File name	Size	Action
base_stations.parquet	12.8 KB	Remove

Select the file and from 3 dots side of more select "properties"



Name	Last Modified	Content Type
base_stations.parquet	1/14/2024, 11:06:29 AM	

Copy the **ABFSS** path



Properties

Name

mydata/base_stations.parquet

URL

https://mydata08.dfs.core.windows.net/myfiles08/mydata/base_stations.parquet

ABFSS Path

abfss://myfiles08@mydata08.dfs.core.windows.net/mydata/base_stations.parquet

Last modified

1/14/2024, 11:06:29 AM

Cache Control

max-age=0

Note: Below is example do not copy the same, copy from your path.

abfss://myfiles08@mydata08.dfs.core.windows.net/mydata/base_stations.parquet

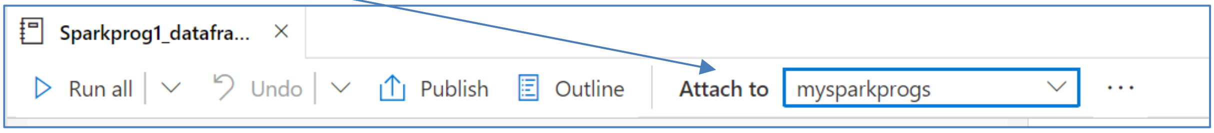
Step 3

Creating dataframe from parquet file.

Go to **Develop Hub** and select **Notebook**.

Give Notebook name as "**Sparkprog1_dataframe**"

Select **Attach to** from top menu and provide your spark pool name.



Attach to mysparkprogs

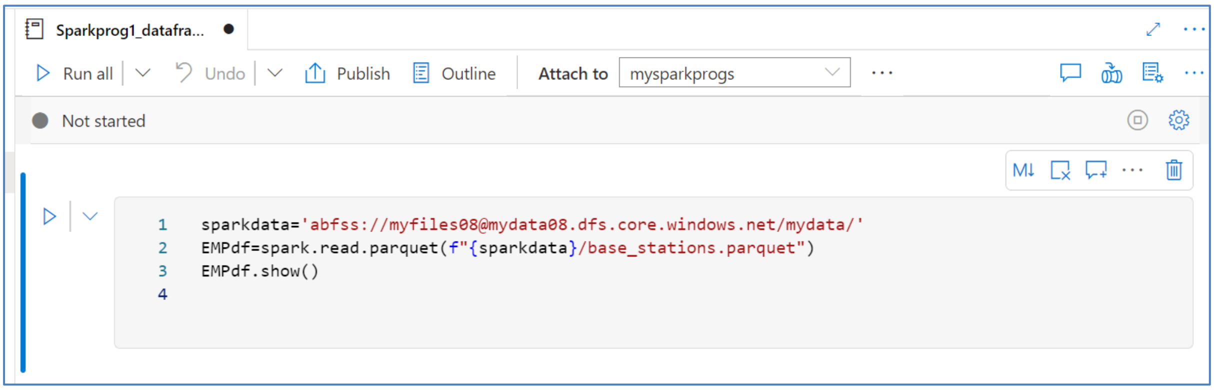
Enter the below under the code.

Note: Please provide your abfss path copied in earlier, under storage folder name.

sparkdata='abfss://myfiles08@mydata08.dfs.core.windows.net/mydata/'

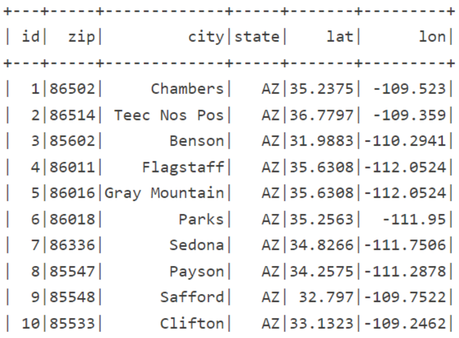
EMPdf=spark.read.parquet(f"{sparkdata}/base_stations.parquet")

EMPdf.show()



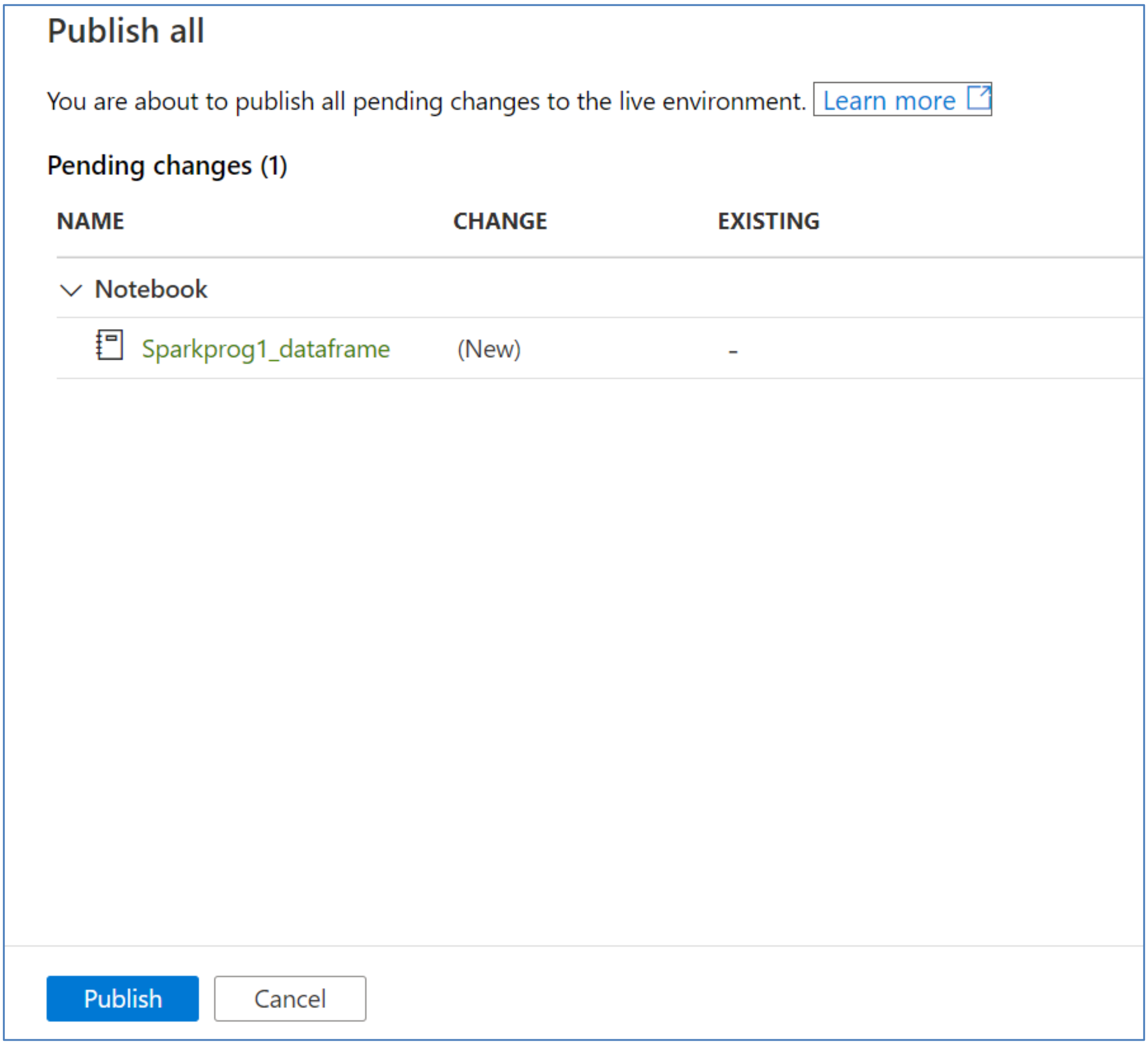
Attach to mysparkprogs

```
sparkdata='abfss://myfiles08@mydata08.dfs.core.windows.net/mydata/'
EMPdf=spark.read.parquet(f"{sparkdata}/base_stations.parquet")
EMPdf.show()
```



```
+---+-----+-------------+-----+-------+---------+
| id|  zip|         city|state|    lat|      lon|
+---+-----+-------------+-----+-------+---------+
|  1|86502|     Chambers|   AZ|35.2375| -109.523|
|  2|86514| Teec Nos Pos|   AZ|36.7797| -109.359|
|  3|85602|       Benson|   AZ|31.9883|-110.2941|
|  4|86011|    Flagstaff|   AZ|35.6308|-112.0524|
|  5|86016|Gray Mountain|   AZ|35.6308|-112.0524|
|  6|86018|        Parks|   AZ|35.2563|  -111.95|
|  7|86336|       Sedona|   AZ|34.8266|-111.7506|
|  8|85547|       Payson|   AZ|34.2575|-111.2878|
|  9|85548|      Safford|   AZ| 32.797|-109.7522|
| 10|85533|      Clifton|   AZ|33.1323|-109.2462|
```

Click on **Publish All** → Publish to save the notebook.



Publish all

You are about to publish all pending changes to the live environment. Learn more

Pending changes (1)

NAME	CHANGE	EXISTING
Notebook		
Sparkprog1_dataframe	(New)	-

*** **Happy Learning** ***