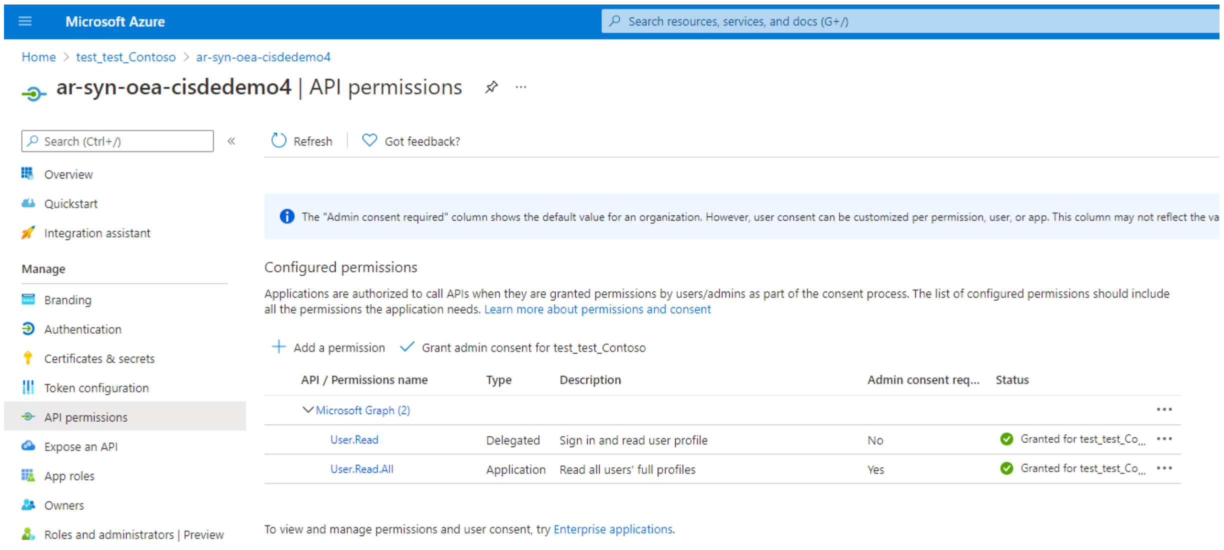
Setting up a pipeline to pull user data from Graph API

1) create linked service to GraphAPI

First go to AAD -> App registrations -> New registration, and create a new app registration specifically for accessing the Graph API from Synapse.

Then click Add Permission, then select Microsoft Graph, then select Application permissions, then choose the User.Read.All permission and click on the Add permissions button.

Now go to certificates & secrets and add a new client secret, and copy the value.


Microsoft Azure | Search resources, services, and docs (G+/)

Home > test_test_Contoso > ar-syn-oea-cisdedemo4

ar-syn-oea-cisdedemo4 | API permissions

Search (Ctrl+/) | Refresh | Got feedback?

Overview · Quickstart · Integration assistant

Manage: Branding · Authentication · Certificates & secrets · Token configuration · API permissions · Expose an API · App roles · Owners · Roles and administrators | Preview

The "Admin consent required" column shows the default value for an organization. However, user consent can be customized per permission, user, or app. This column may not reflect the va

Configured permissions

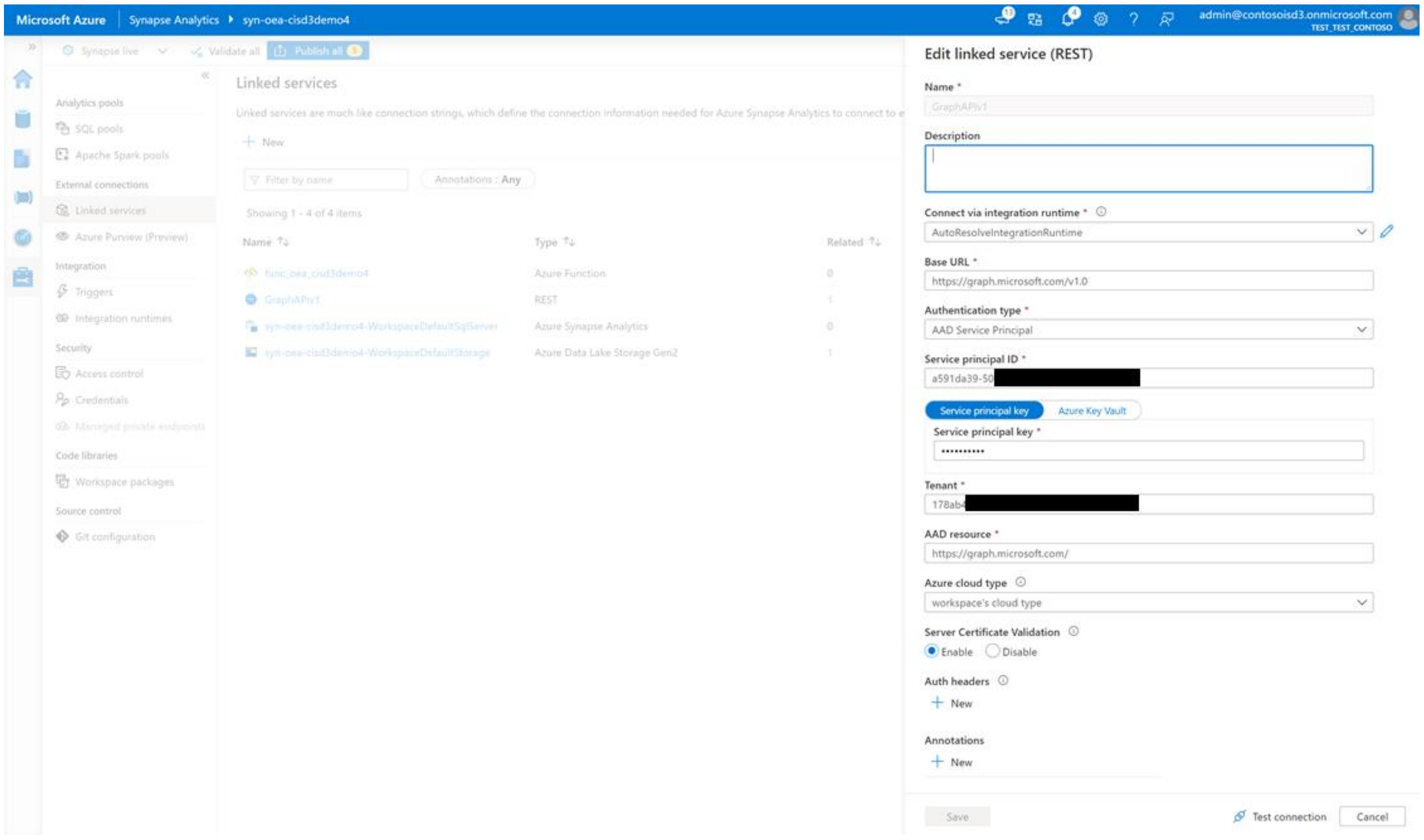
Applications are authorized to call APIs when they are granted permissions by users/admins as part of the consent process. The list of configured permissions should include all the permissions the application needs. Learn more about permissions and consent

+ Add a permission ✓ Grant admin consent for test_test_Contoso

API / Permissions name	Type	Description	Admin consent req...	Status
Microsoft Graph (2)				
User.Read	Delegated	Sign in and read user profile	No	✓ Granted for test_test_Co...
User.Read.All	Application	Read all users' full profiles	Yes	✓ Granted for test_test_Co...

To view and manage permissions and user consent, try Enterprise applications.

Now in synapse studio, go to linked services and add a REST service for Graph API and select AAD Service Principal as the Authentication type.


Microsoft Azure | Synapse Analytics ▸ syn-oea-cisd3demo4 | admin@contosoisd3.onmicrosoft.com TEST_TEST_CONTOSO

Synapse live | Validate all | Publish all

Analytics pools: SQL pools · Apache Spark pools
External connections: Linked services · Azure Purview (Preview)
Integration: Triggers · Integration runtimes
Security: Access control · Credentials · Managed private endpoints
Code libraries: Workspace packages
Source control: Git configuration

Linked services

Linked services are much like connection strings, which define the connection information needed for Azure Synapse Analytics to connect to e

+ New | Filter by name | Annotations : Any

Showing 1 - 4 of 4 items

Name	Type	Related
func_oea_cisd3demo4	Azure Function	0
GraphAPIv1	REST	1
syn-oea-cisd3demo4-WorkspaceDefaultSqlServer	Azure Synapse Analytics	0
syn-oea-cisd3demo4-WorkspaceDefaultStorage	Azure Data Lake Storage Gen2	1

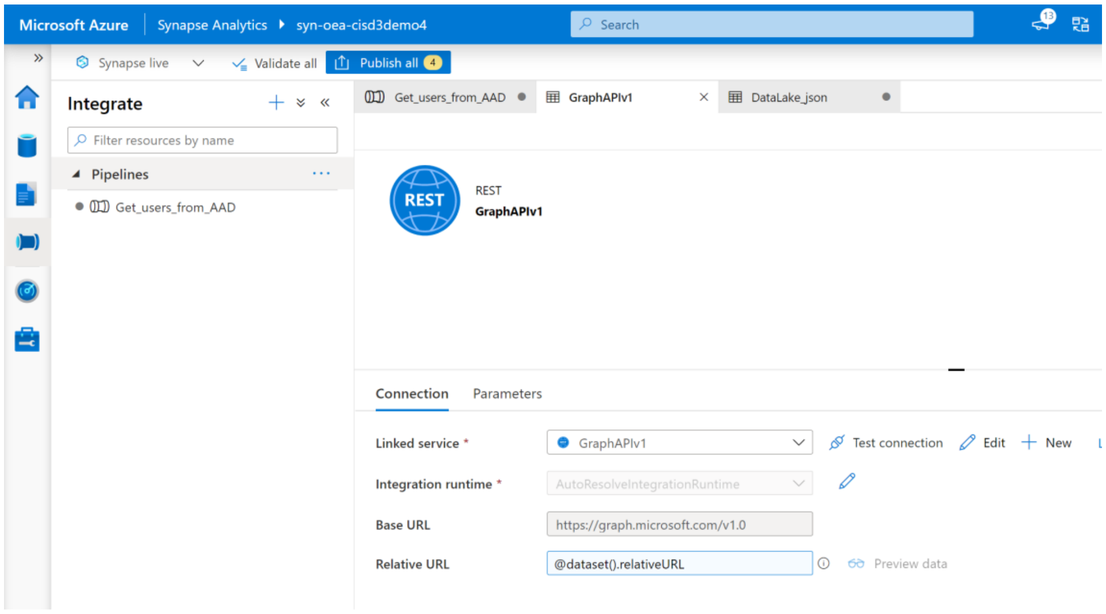
Edit linked service (REST)

Name *	GraphAPIv1
Description	
Connect via integration runtime *	AutoResolveIntegrationRuntime
Base URL *	https://graph.microsoft.com/v1.0
Authentication type *	AAD Service Principal
Service principal ID *	a591da39-50[redacted]
	[✓ Service principal key] Azure Key Vault
Service principal key *	••••••••••
Tenant *	178ab[redacted]
AAD resource *	https://graph.microsoft.com/
Azure cloud type	workspace's cloud type
Server Certificate Validation	[✓ Enable] Disable
Auth headers	+ New
Annotations	+ New

Save | Test connection | Cancel

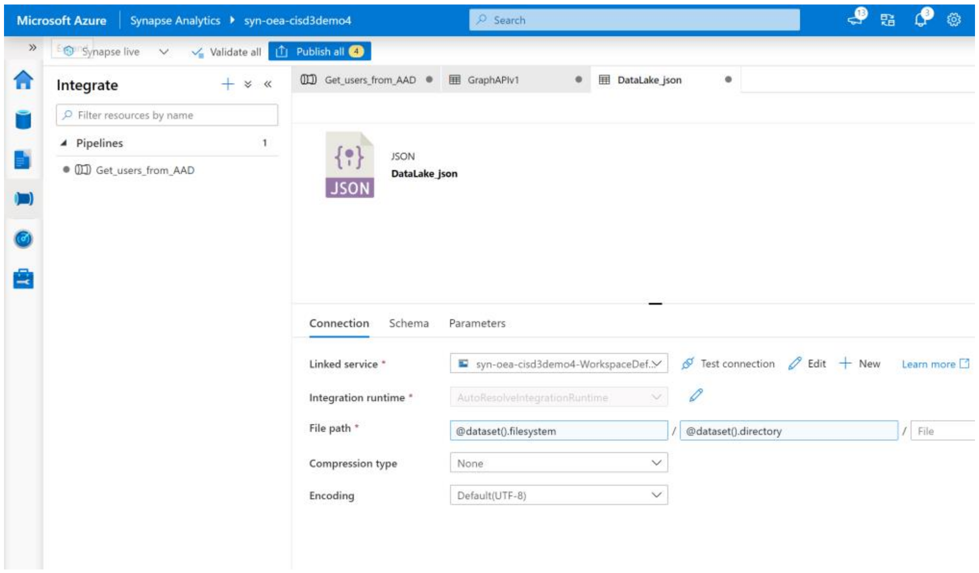
Under service principal key, paste the value of the secret you copied in the previous step.
Under Service principal ID enter the "Application (client) ID" of the app registration created in the previous step (go to the Overview section of the app registration).

2) create a pipeline, and then create a new REST dataset as the source, referring to the GraphAPIv1 linked service

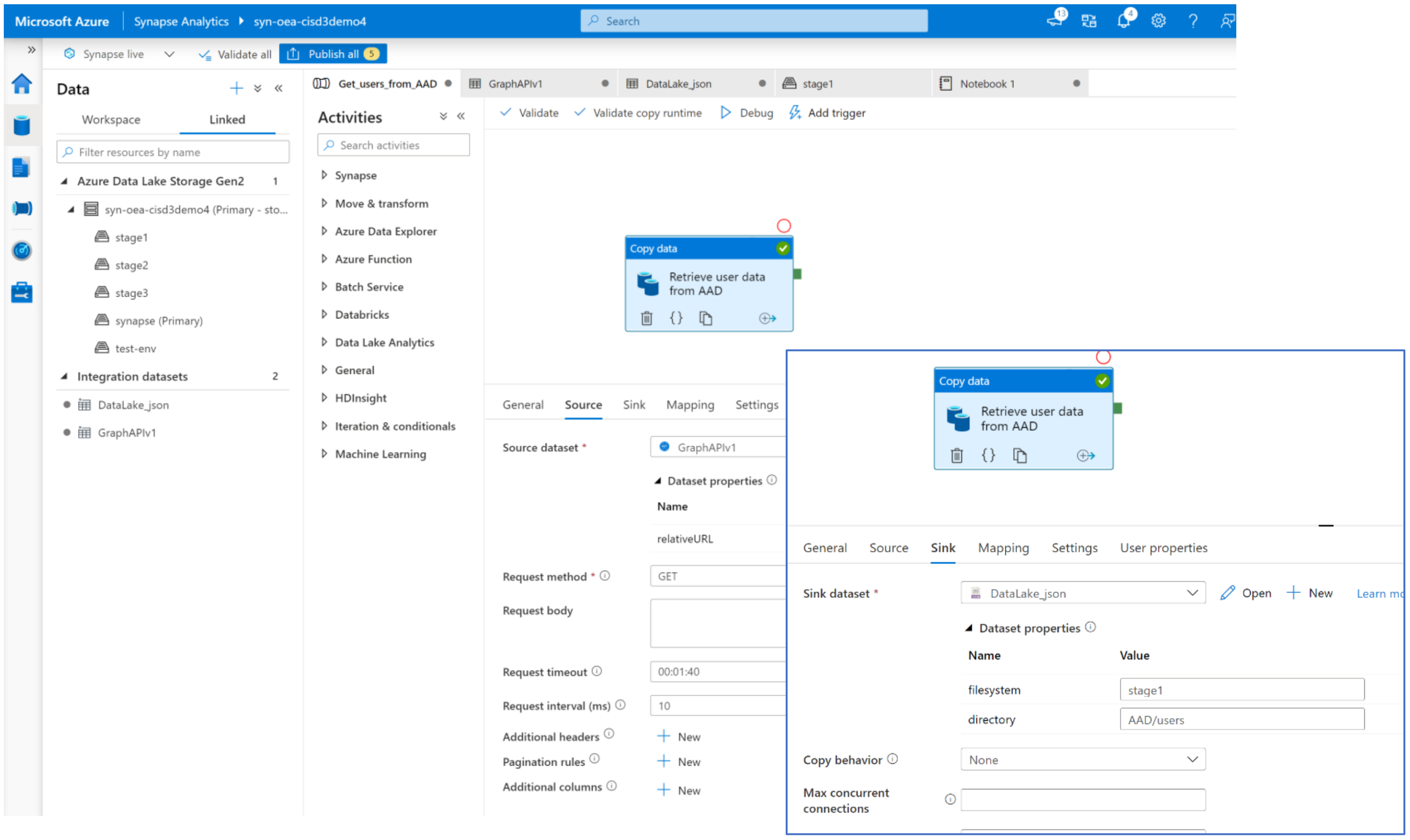


Linked service * GraphAPIv1
Integration runtime * AutoResolveIntegrationRuntime
Base URL https://graph.microsoft.com/v1.0
Relative URL @dataset().relativeURL

3) create a sink dataset going to the data lake



Linked service * syn-oea-cisd3demo4-WorkspaceDef...
Integration runtime * AutoResolveIntegrationRuntime
File path * @dataset().filesystem / @dataset().directory / File
Compression type None
Encoding Default(UTF-8)



Azure Data Lake Storage Gen2 1
- syn-oea-cisd3demo4 (Primary - sto...
  - stage1
  - stage2
  - stage3
  - synapse (Primary)
  - test-env

Integration datasets 2
- DataLake_json
- GraphAPIv1

Source tab:

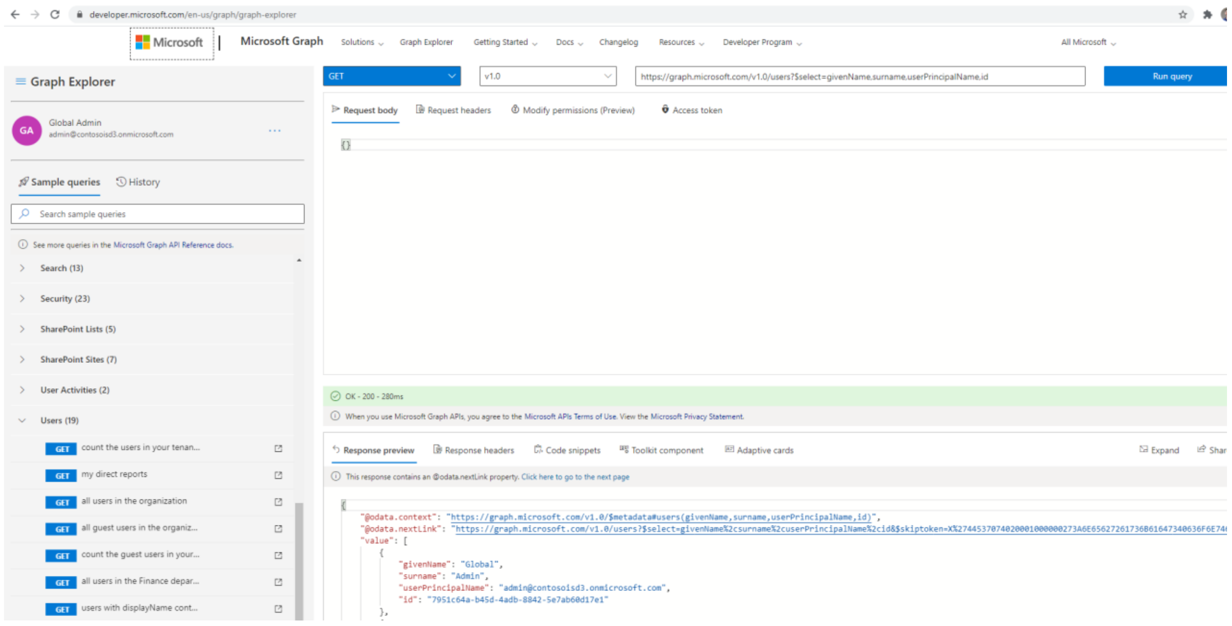
Source dataset *	GraphAPIv1
Dataset properties: Name	relativeURL
Request method *	GET
Request body	
Request timeout	00:01:40
Request interval (ms)	10
Additional headers	+ New
Pagination rules	+ New
Additional columns	+ New

Sink tab:

Sink dataset *	DataLake_json	
Dataset properties	Name	Value
	filesystem	stage1
	directory	AAD/users
Copy behavior	None	
Max concurrent connections		

You can reduce the data returned by selecting specific attributes in the relativeURL, like this: users?$select=givenName,surname,userPrincipalName,id

Use the Graph Explorer to try it out:



GET v1.0 https://graph.microsoft.com/v1.0/users?$select=givenName,surname,userPrincipalName,id

OK - 200 - 280ms

This response contains an @odata.nextLink property. Click here to go to the next page

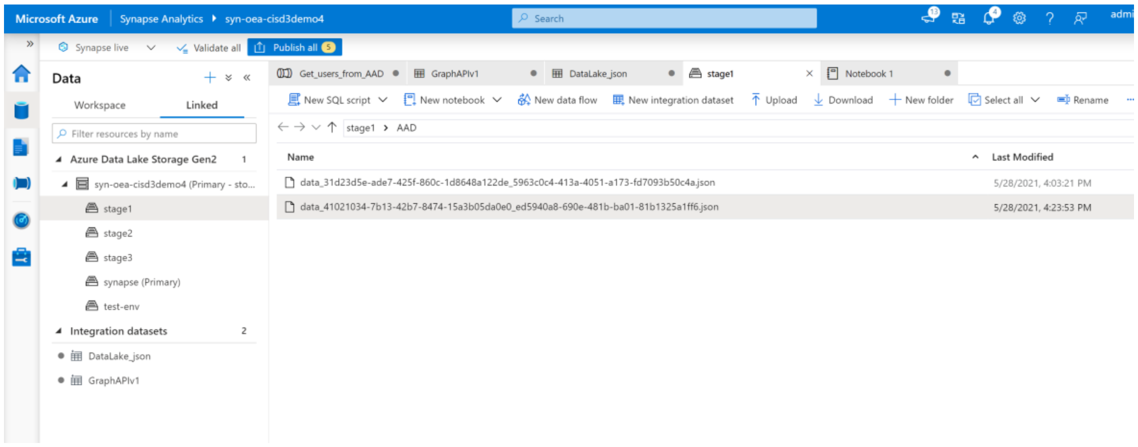
```
{
    "@odata.context": "https://graph.microsoft.com/v1.0/$metadata#users(givenName,surname,userPrincipalName,id)",
    "@odata.nextLink": "https://graph.microsoft.com/v1.0/users?$select=givenName%2csurname%2cuserPrincipalName%2cid&$skiptoken=X%2744537074020001000000273A6E65627261736B61647340636F6E74
    "value": [
        {
            "givenName": "Global",
            "surname": "Admin",
            "userPrincipalName": "admin@contosoisd3.onmicrosoft.com",
            "id": "7951c64a-b45d-4adb-8842-5e7ab60d17e1"
        },
```

If you run the pipeline and get a failure because of lack of access like this:

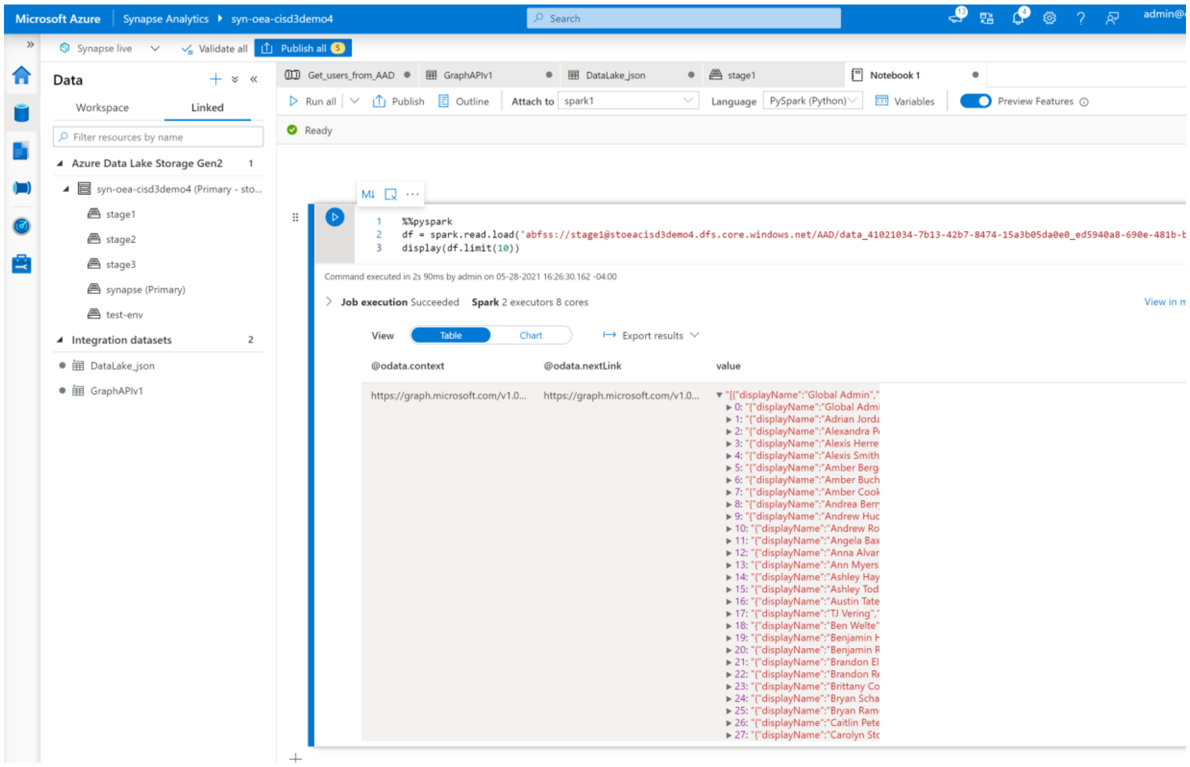
```
{ "errorCode": "2200", "message": "Failure happened on 'Source' side.
ErrorCode=RestSourceCallFailed,'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Messa ge=The
HttpStatusCode 403 indicates failure.\nRequest URL: https://graph.microsoft.com/v1.0/users?$top=3\nResponse
payload:{\"error\":{\"code\":\"Authorization_RequestDenied\",\"message\":\"Insufficient privileges to complete the
operation.\",\"innerError\":{\"date\":\"2021-05-28T19:31:00\",\"request-id\":\"abc221e5-a4d1-4845-8a5d-
c26353b99af3\",\"client-request-id\":\"abc221e5-a4d1-4845-8a5d-
c26353b99af3\"}}},Source=Microsoft.DataTransfer.ClientLibrary,'", "failureType": "UserError", "target": "Copy data1",
"details": [] }
```

you'll need to double check the app registration you created and the API permissions assigned to it.

You can then go into Synapse and navigate to your data lake and see the json file. Right click on it and select Load to dataframe, then run the Synapse notebook.



Name	Last Modified
data_31d23d5e-ade7-425f-860c-1d8648a122de_5963c0c4-413a-4051-a173-fd7093b50c4a.json	5/28/2021, 4:03:21 PM
data_41021034-7b13-42b7-8474-15a3b05da0e0_ed5940a8-690e-481b-ba01-81b1325a1ff6.json	5/28/2021, 4:23:53 PM



```
%%pyspark
df = spark.read.load('abfss://stage1@stoeacisd3demo4.dfs.core.windows.net/AAD/data_41021034-7b13-42b7-8474-15a3b05da0e0_ed5940a8-690e-481b-b
display(df.limit(10))
```

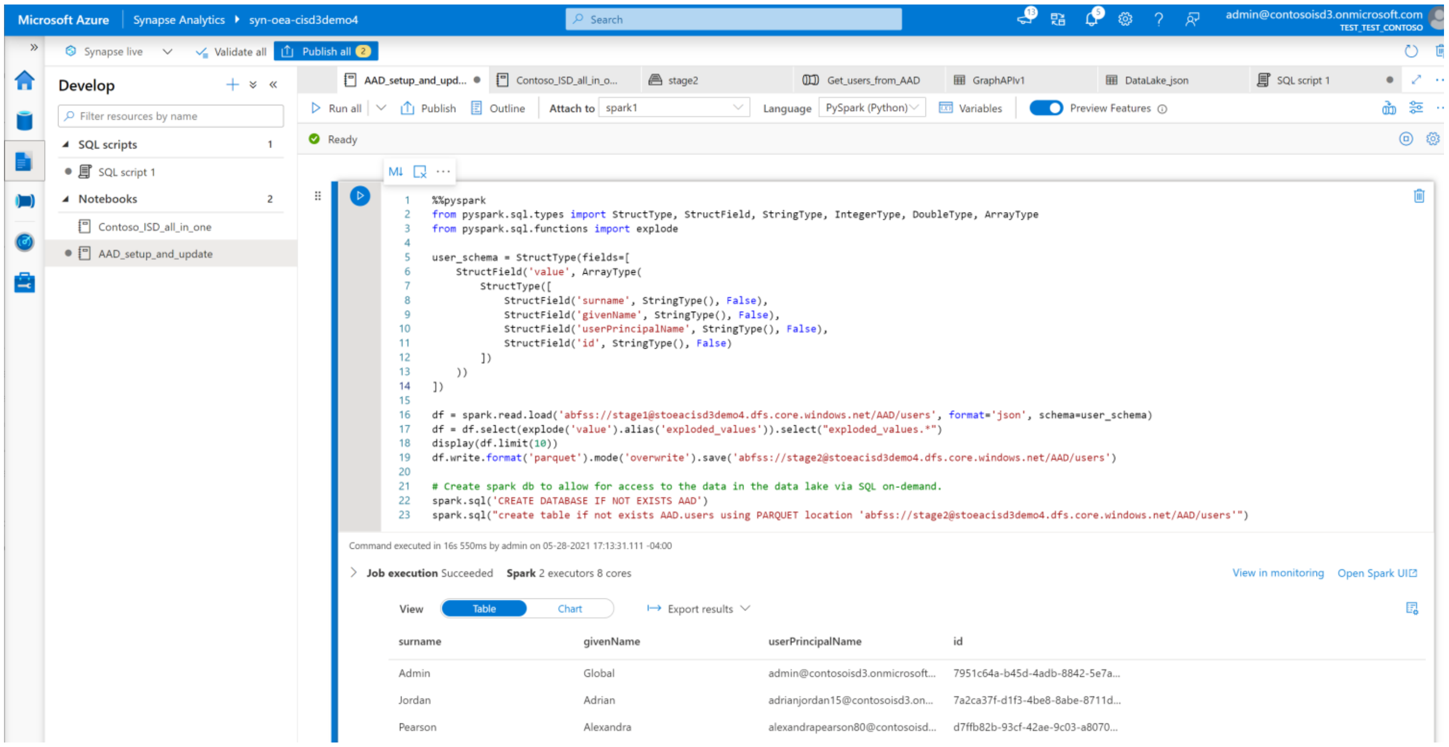
```
Command executed in 2s 90ms by admin on 05-28-2021 16:26:30.162 -04:00

Job execution Succeeded   Spark 2 executors 8 cores
```

@odata.context	@odata.nextLink	value
https://graph.microsoft.com/v1.0...	https://graph.microsoft.com/v1.0...	"[{"displayName":"Global Admin","
		0: "{"displayName":"Global Admi
		1: "{"displayName":"Adrian Jorda
		2: "{"displayName":"Alexandra P
		3: "{"displayName":"Alexis Herre
		4: "{"displayName":"Alexis Smith
		5: "{"displayName":"Amber Berg
		6: "{"displayName":"Amber Buch
		7: "{"displayName":"Amber Cook
		8: "{"displayName":"Andrea Berr
		9: "{"displayName":"Andrew Huc
		10: "{"displayName":"Andrew Ro
		11: "{"displayName":"Angela Bax
		12: "{"displayName":"Anna Alvar
		13: "{"displayName":"Ann Myers
		14: "{"displayName":"Ashley Hay
		15: "{"displayName":"Ashley Tod
		16: "{"displayName":"Austin Tate
		17: "{"displayName":"TJ Vering","
		18: "{"displayName":"Ben Welte"
		19: "{"displayName":"Benjamin H
		20: "{"displayName":"Benjamin R
		21: "{"displayName":"Brandon El
		22: "{"displayName":"Brandon R
		23: "{"displayName":"Brittany Co
		24: "{"displayName":"Bryan Scha
		25: "{"displayName":"Bryan Ram
		26: "{"displayName":"Caitlin Pete
		27: "{"displayName":"Carolyn Sto

Now you need to process this user data and write it to stage 2 in a more usable way, and also create a spark db that can be easily queried from Power BI.

Run this notebook to do that:



```
%%pyspark
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, DoubleType, ArrayType
from pyspark.sql.functions import explode

user_schema = StructType(fields=[
    StructField('value', ArrayType(
        StructType([
            StructField('surname', StringType(), False),
            StructField('givenName', StringType(), False),
            StructField('userPrincipalName', StringType(), False),
            StructField('id', StringType(), False)
        ])
    ))
])

df = spark.read.load('abfss://stage1@stoeacisd3demo4.dfs.core.windows.net/AAD/users', format='json', schema=user_schema)
df = df.select(explode('value').alias('exploded_values')).select("exploded_values.*")
display(df.limit(10))
df.write.format('parquet').mode('overwrite').save('abfss://stage2@stoeacisd3demo4.dfs.core.windows.net/AAD/users')

# Create spark db to allow for access to the data in the data lake via SQL on-demand.
spark.sql('CREATE DATABASE IF NOT EXISTS AAD')
spark.sql("create table if not exists AAD.users using PARQUET location 'abfss://stage2@stoeacisd3demo4.dfs.core.windows.net/AAD/users'")
```

Command executed in 16s 550ms by admin on 05-28-2021 17:13:31.111 -04:00

Job execution Succeeded Spark 2 executors 8 cores

surname	givenName	userPrincipalName	id
Admin	Global	admin@contosoisd3.onmicrosoft...	7951c64a-b45d-4adb-8842-5e7a...
Jordan	Adrian	adrianjordan15@contosoisd3.on...	7a2ca37f-d1f3-4be8-8abe-8711d...
Pearson	Alexandra	alexandrapearson80@contosoisd...	d7ffb82b-93cf-42ae-9c03-a8070...

Here’s the script (you have to change the url values to use your storage account name):

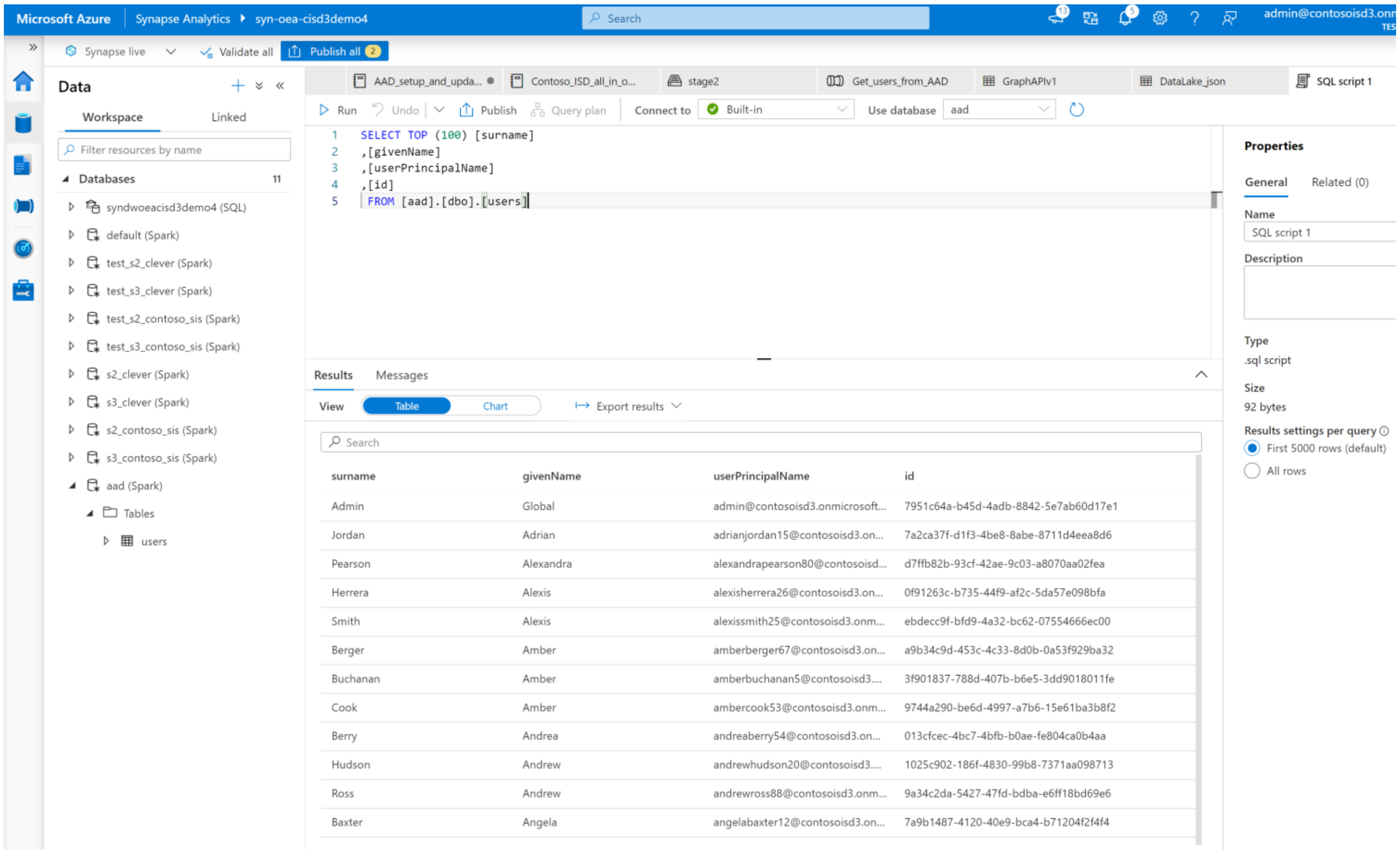
```
%%pyspark
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, DoubleType, ArrayType
from pyspark.sql.functions import explode

user_schema = StructType(fields=[
    StructField('value', ArrayType(
        StructType([
            StructField('surname', StringType(), False),
            StructField('givenName', StringType(), False),
            StructField('userPrincipalName', StringType(), False),
            StructField('id', StringType(), False)
        ])
    ))
])

df = spark.read.load('abfss://stage1@stoeacisd3demo4.dfs.core.windows.net/AAD/users', format='json', schema=user_schema)
df = df.select(explode('value').alias('exploded_values')).select("exploded_values.*")
display(df.limit(10))
df.write.format('parquet').mode('overwrite').save('abfss://stage2@stoeacisd3demo4.dfs.core.windows.net/AAD/users')

# Create spark db to allow for access to the data in the data lake via SQL on-demand.
spark.sql('CREATE DATABASE IF NOT EXISTS AAD')
spark.sql("create table if not exists AAD.users using PARQUET location 'abfss://stage2@stoeacisd3demo4.dfs.core.windows.ne
t/AAD/users'")
```

Now you’re able to query the user data from the spark db:



Name: SQL script 1

Results settings per query: [✓ First 5000 rows (default)] All rows

```
SELECT TOP (100) [surname]
,[givenName]
,[userPrincipalName]
,[id]
 FROM [aad].[dbo].[users]
```

surname	givenName	userPrincipalName	id
Admin	Global	admin@contosoisd3.onmicrosoft...	7951c64a-b45d-4adb-8842-5e7ab60d17e1
Jordan	Adrian	adrianjordan15@contosoisd3.on...	7a2ca37f-d1f3-4be8-8abe-8711d4eea8d6
Pearson	Alexandra	alexandrapearson80@contosoisd...	d7ffb82b-93cf-42ae-9c03-a8070aa02fea
Herrera	Alexis	alexisherrera26@contosoisd3.on...	0f91263c-b735-44f9-af2c-5da57e098bfa
Smith	Alexis	alexissmith25@contosoisd3.onm...	ebdecc9f-bfd9-4a32-bc62-07554666ec00
Berger	Amber	amberberger67@contosoisd3.on...	a9b34c9d-453c-4c33-8d0b-0a53f929ba32
Buchanan	Amber	amberbuchanan5@contosoisd3....	3f901837-788d-407b-b6e5-3dd9018011fe
Cook	Amber	ambercook53@contosoisd3.onm...	9744a290-be6d-4997-a7b6-15e61ba3b8f2
Berry	Andrea	andreaberry54@contosoisd3.on...	013cfcec-4bc7-4bfb-b0ae-fe804ca0b4aa
Hudson	Andrew	andrewhudson20@contosoisd3....	1025c902-186f-4830-99b8-7371aa098713
Ross	Andrew	andrewross88@contosoisd3.onm...	9a34c2da-5427-47fd-bdba-e6ff18bd69e6
Baxter	Angela	angelabaxter12@contosoisd3.on...	7a9b1487-4120-40e9-bca4-b71204f2f4f4