Data Engineering in the Cloud

ETL Using Apache Spark

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

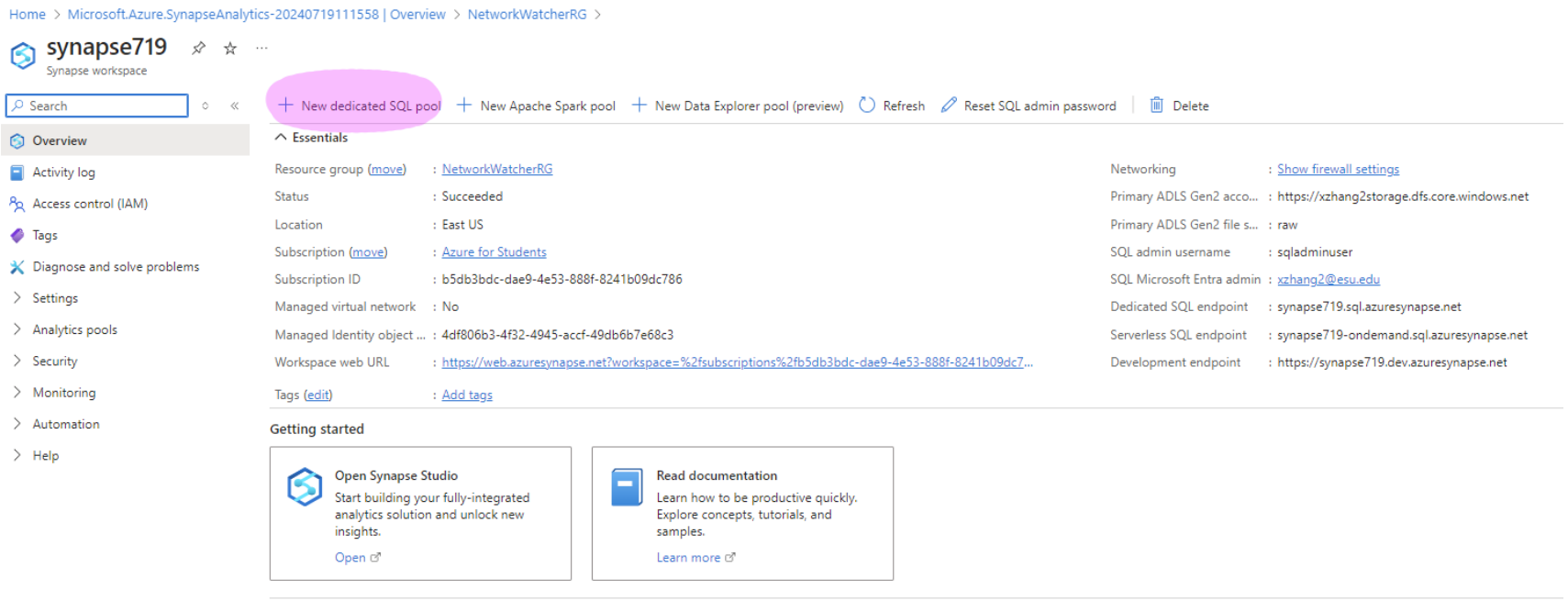
- Prerequisite
- Data exploration in Synapse Studio

Prerequisite

- Createa Data Lake Storage account if you do not have one (Hierarchical namspace enabled)
  - A resource group is needed
- Create an Azure Synapse workspace (Lab 2 in Lecture04)
  - Once the deployment is completed, click on the `go to resource` button

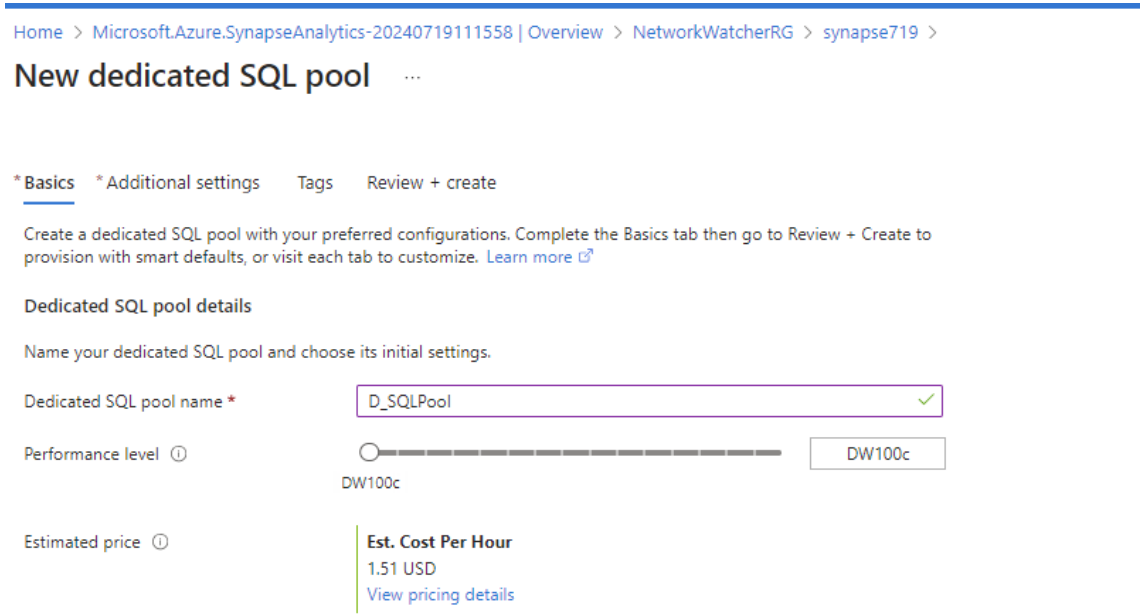
Prerequisite

- Now open the synapse workspace, click on the `New dedicated SQL pool` button



Prerequisite

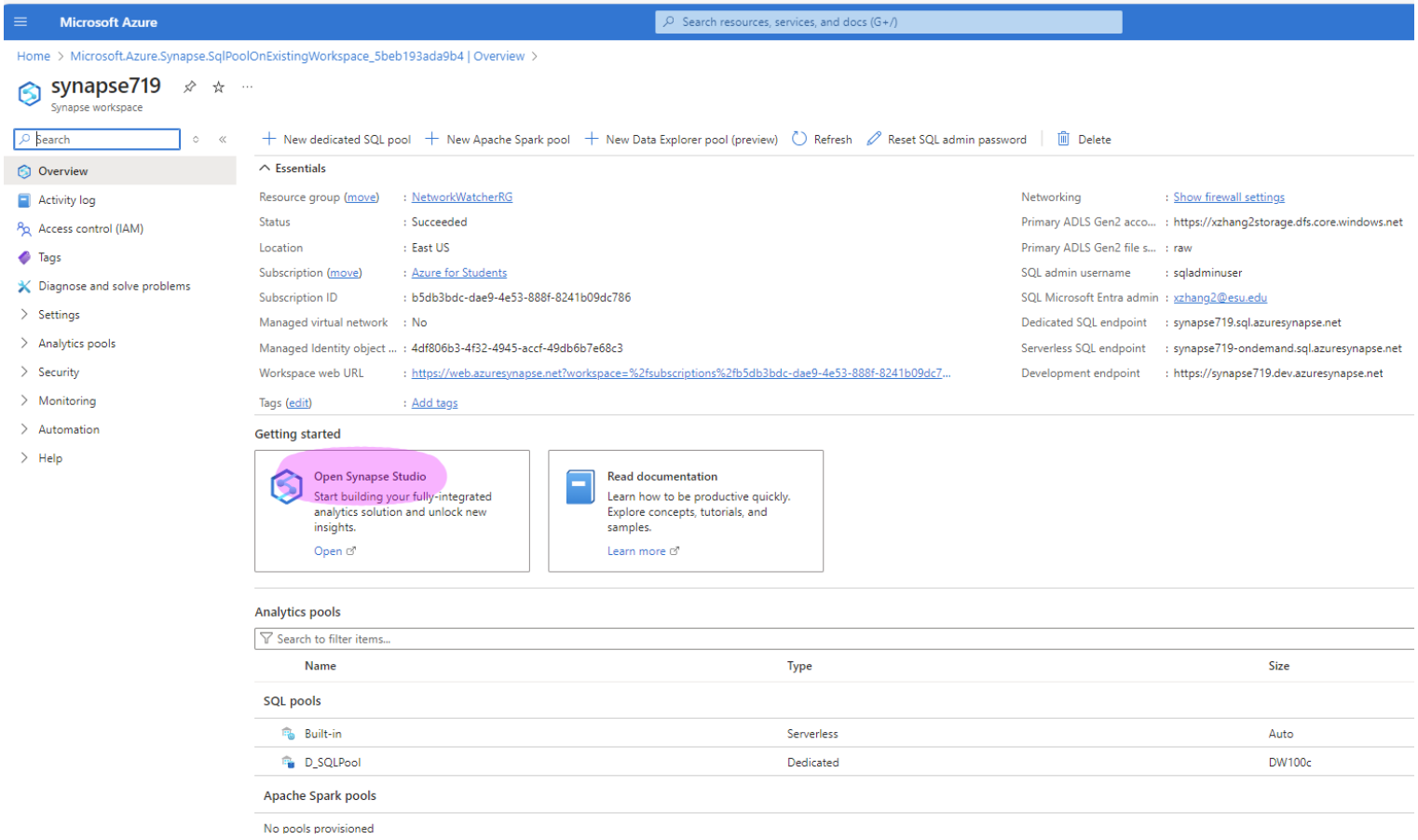
- Specify the name for the SQL pool.
  - Choose the performance level **DW100c**



- Click on the `review + create` button and then click on the `Create` button once the configuration is verified and the SQL pool will be configured successfully

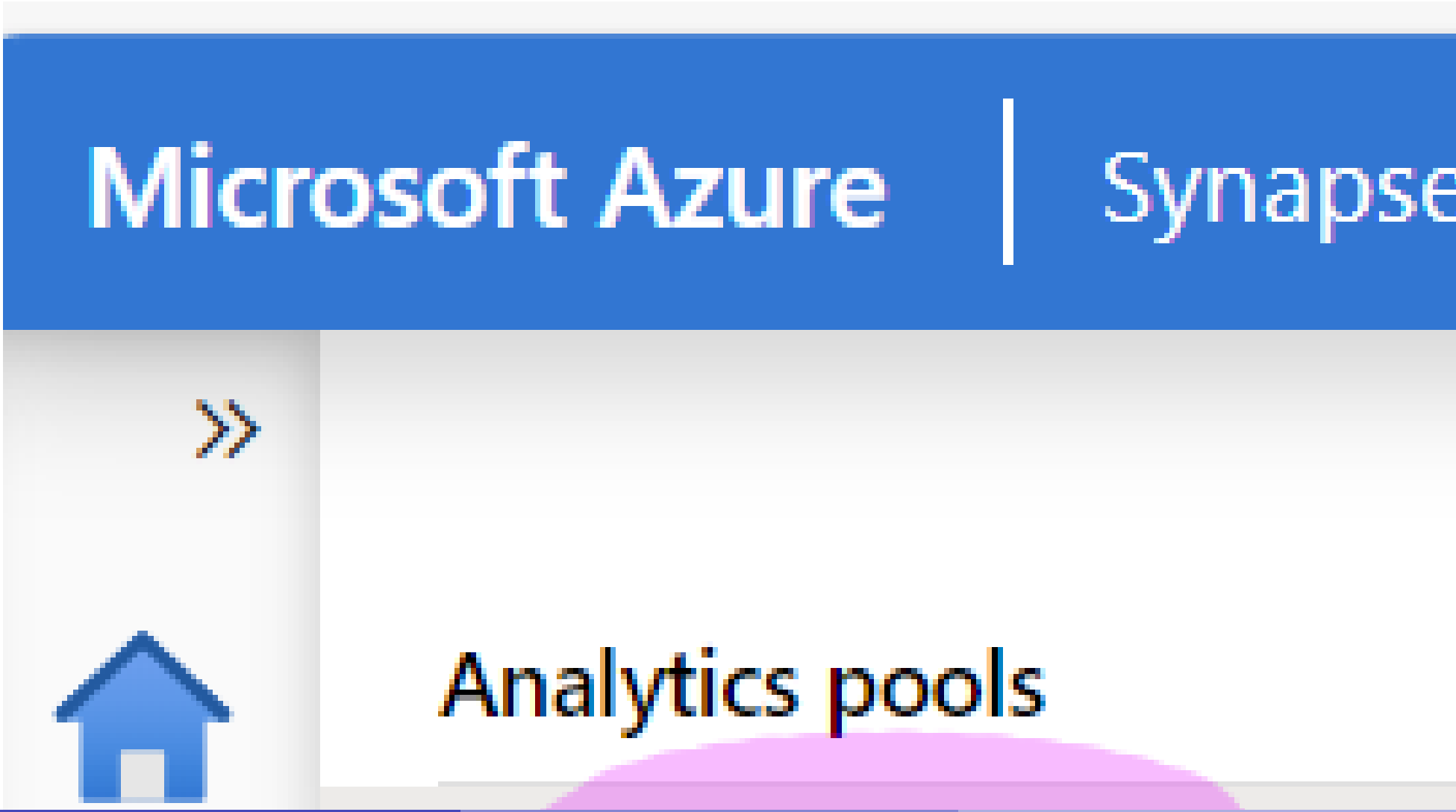
Prerequisite

- Open the `synapse studio`



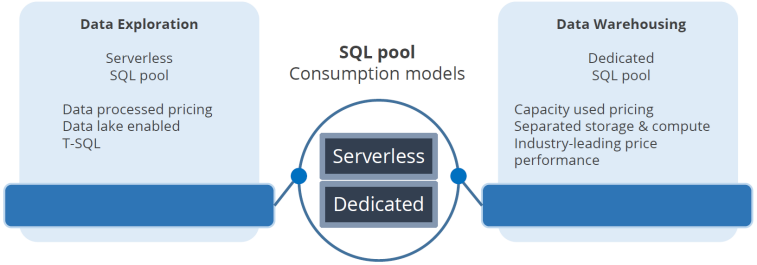
Prerequisite

- And you must be able to see the SQL pool which means it is successfully created



Data exploration in Synapse Studio

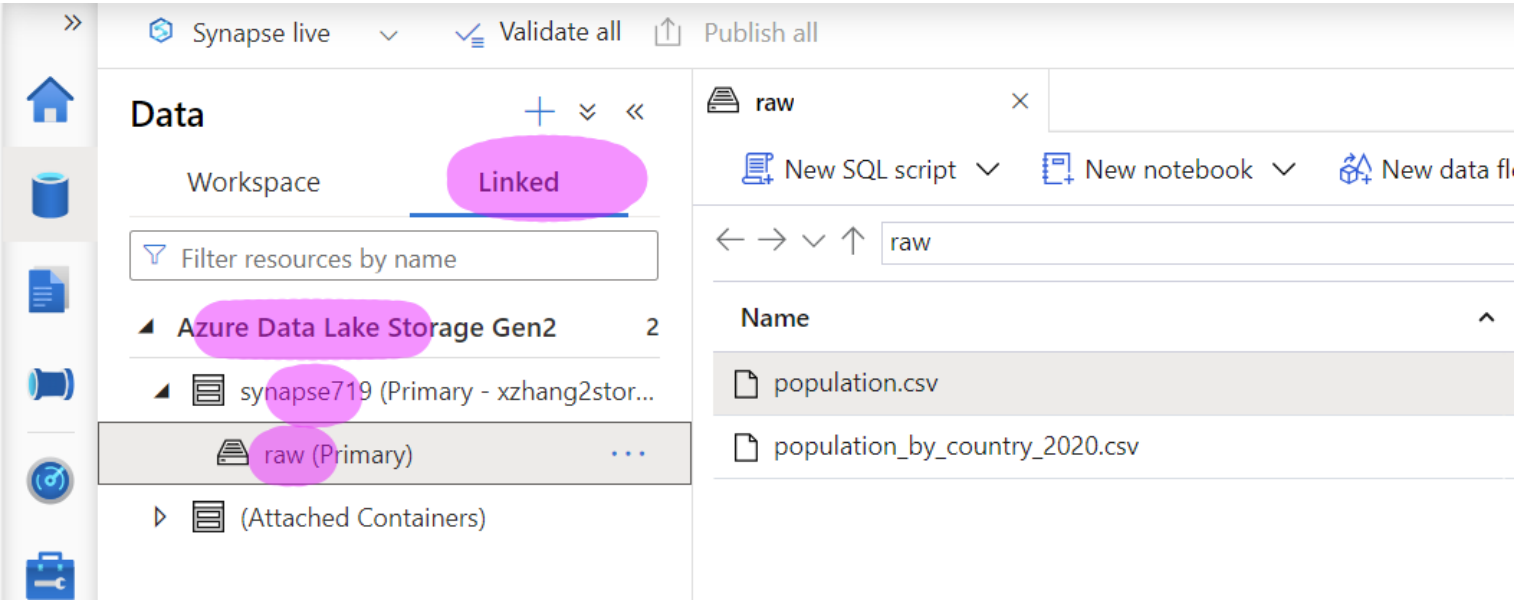
- Data Exploration is the first step for analyzing data before data visualization and statistical analysis.
- Data Integration, Big Data, and Enterprise Data Warehousing are all combined in Azure Synapse Analytics to provide end to end analytics at cloud scale.



- A **synapse notebook** is an interface used to create a file consisting of live code and text by validating ideas and fetch insights from the data.

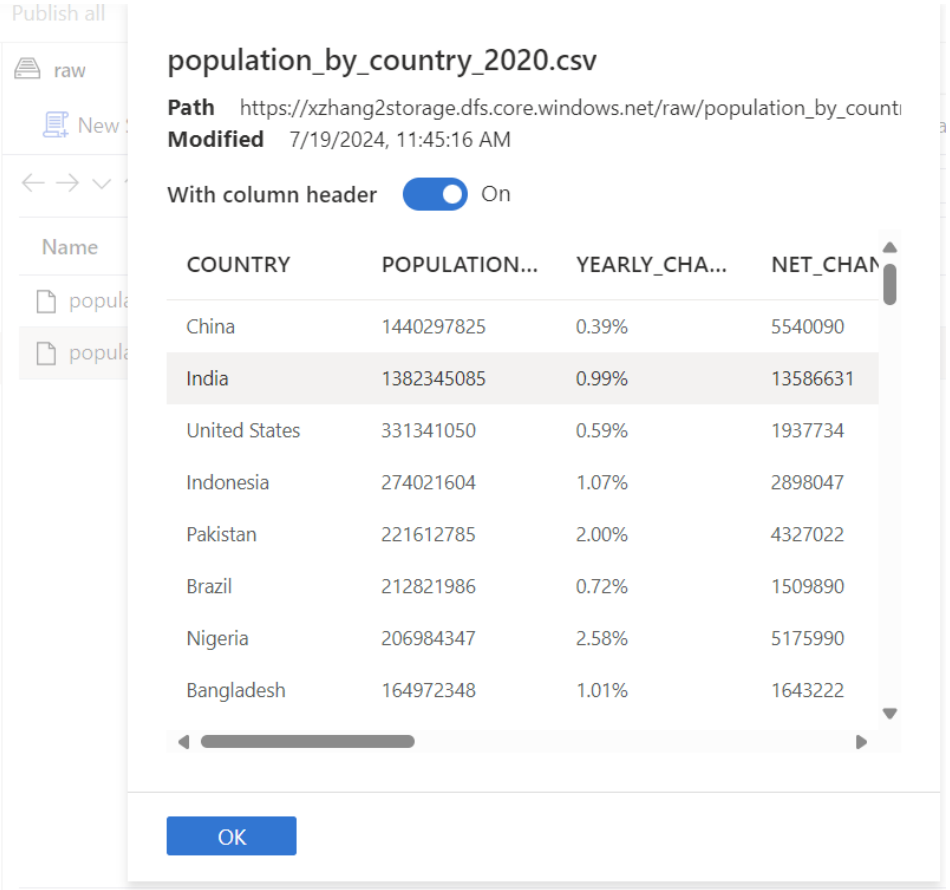
Data exploration in Synapse Studio

- Browse the `Data` tab on the left. Click on `Linked` services, expand the storage account, and click on the container linked to the Synapse.



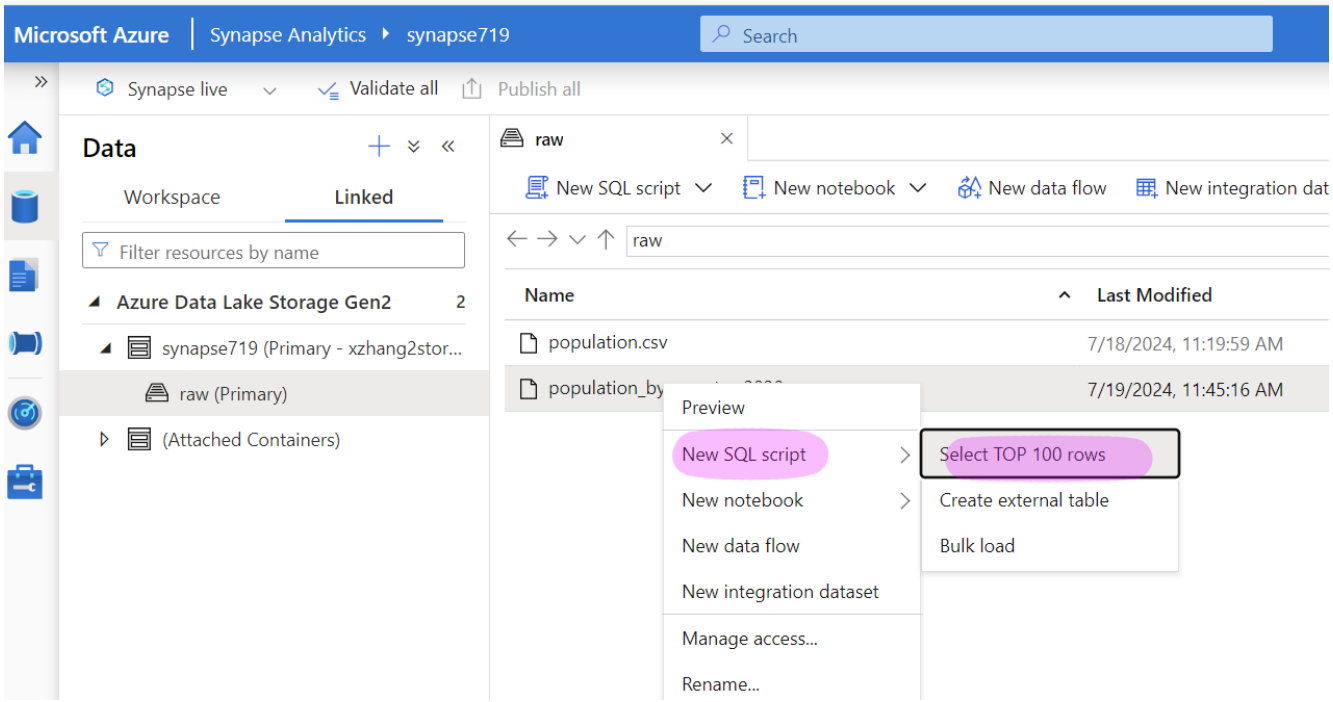
Data exploration in Synapse Studio

- Now, you will be able to see the CSV file `population.csv` that you have uploaded inside the container. Go to your CSV file, right-click, and click on the `preview` option
- You will be able to see and explore the data. The `Preview` feature in Synapse Studio allows you to quickly explore the contents of a file without writing any code. This is a good method to obtain basic knowledge of the characteristics and types of data that are present in a single file.



Data exploration in Synapse Studio

- Now we will use the serverless SQL pool to explore the data.
  - Again, go back to the same window, right click on the csv file" and select `New SQL Script` and then select `Select TOP 100 rows`



Data exploration in Synapse Studio

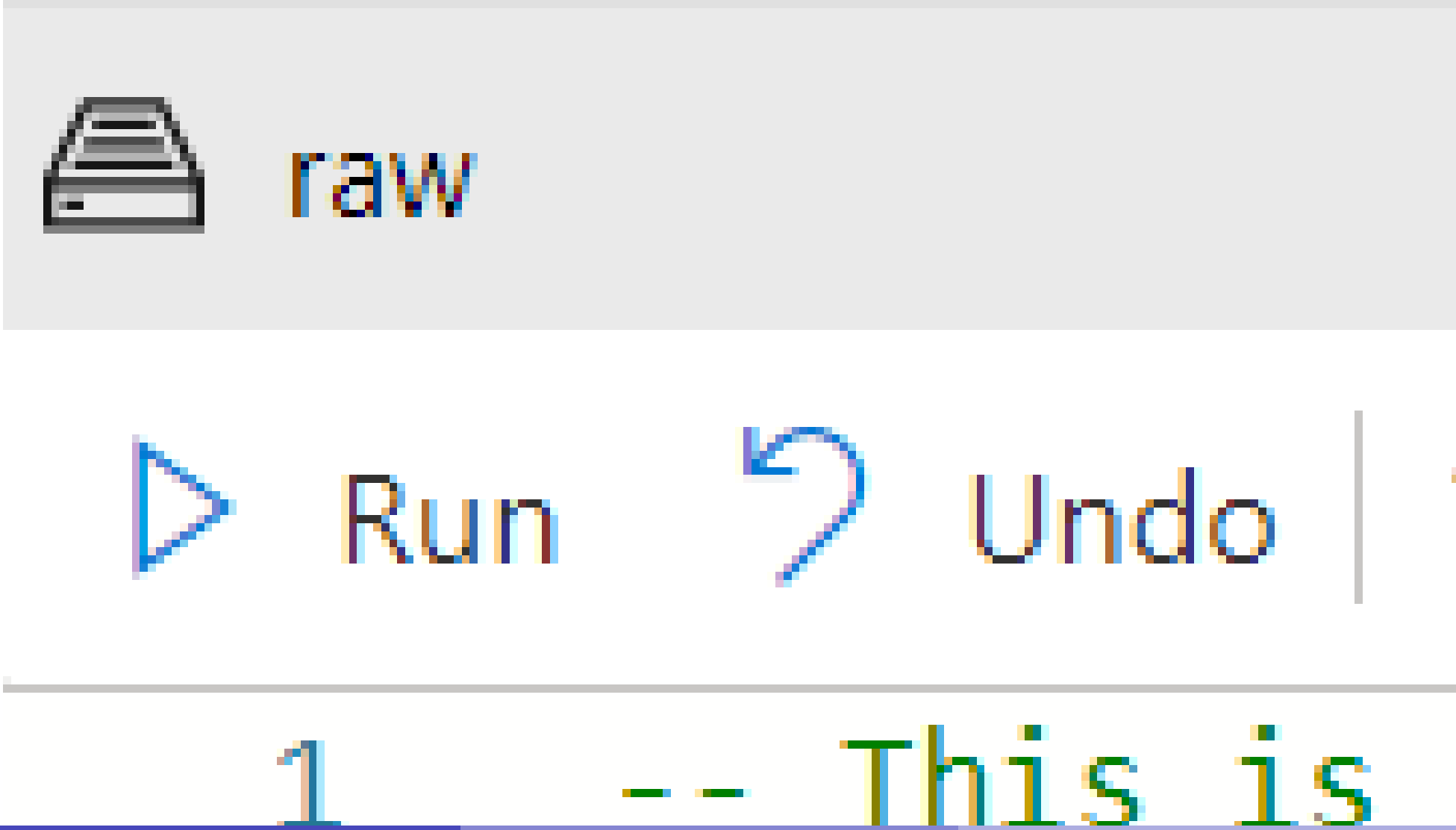
raw | SQL script 1

Run | Undo | Publish | Query plan | Connect to | Built-in | Use dat

```
1  -- This is auto-generated code
2  SELECT
3      TOP 100 *
4  FROM
5      OPENROWSET(
6          BULK 'https://xzhang2storage.dfs.core.windows.net/raw/population_by_count
7          FORMAT = 'CSV',
8          PARSER_VERSION = '2.0'
9      ) AS [result]
10
```

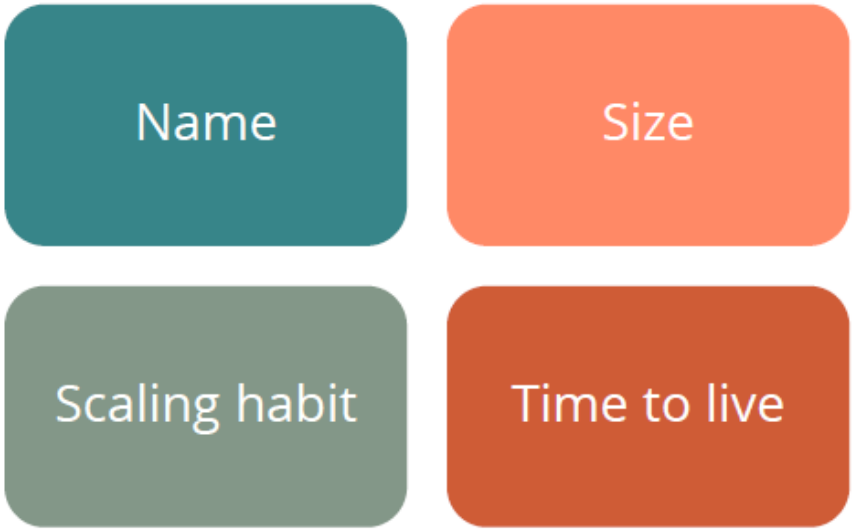
Data exploration in Synapse Studio

- Run the script and you will be able to see the data in the `Results` tab.



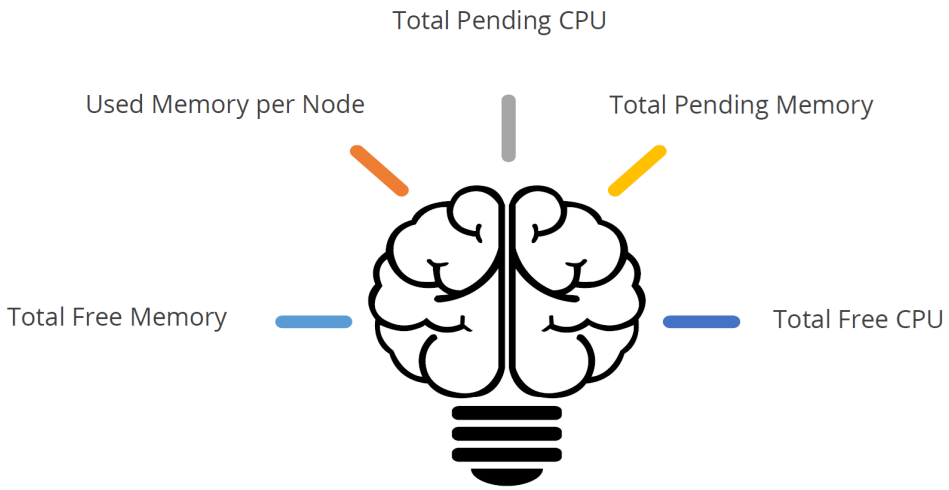
Apache Spark in Azure Synapse Analytics

- **Spark Pool** is a service used in Spark creation for data processing and represented as metadata **without any resources being consumed, executed, or billed**.
- **Spark instance group** is an installation of Apache Spark that helps to run core services like Spark master, shuffle, history, and notebooks.
- Properties to control Spark instance's characteristics:



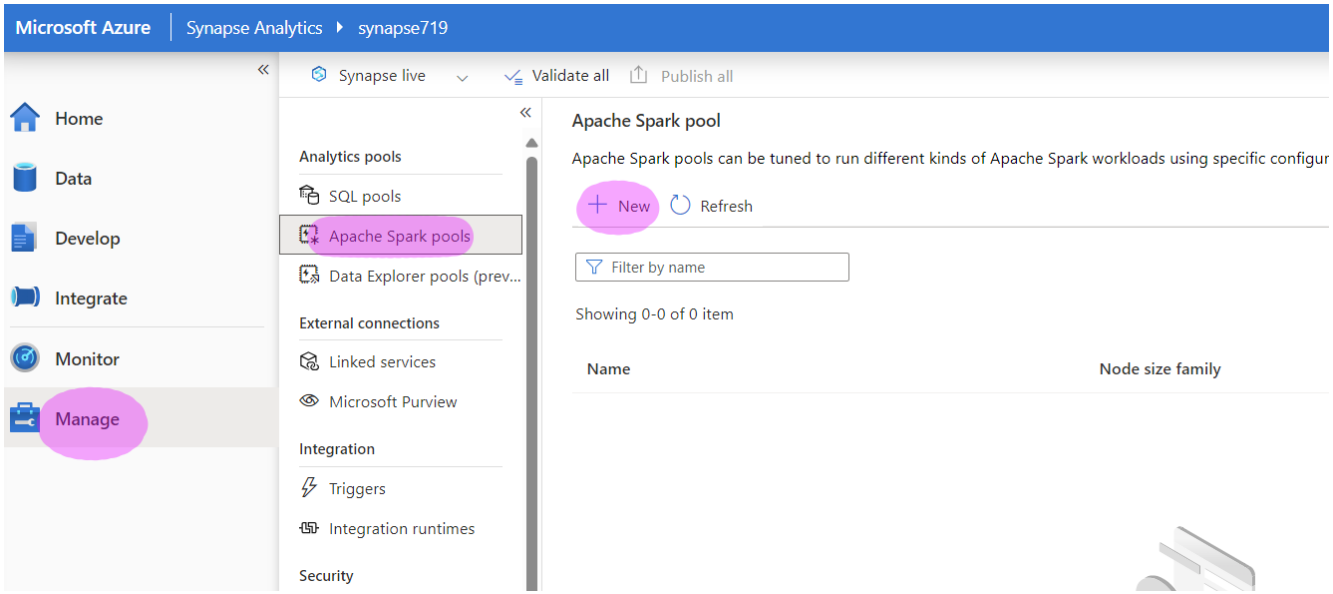
Apache Spark in Azure Synapse Analytics

- Apache Spark Pool Auto Scaling scaling feature is used to automatically increase or decrease the number of nodes on a cluster.
- Apache Spark Pool auto scaling can be built only if the minimum or maximum number of nodes are defined.
- Following are the metrics which auto scale continuously monitors:
  - Total Pending CPU defines the total number of cores that are required to begin the execution of the nodes.



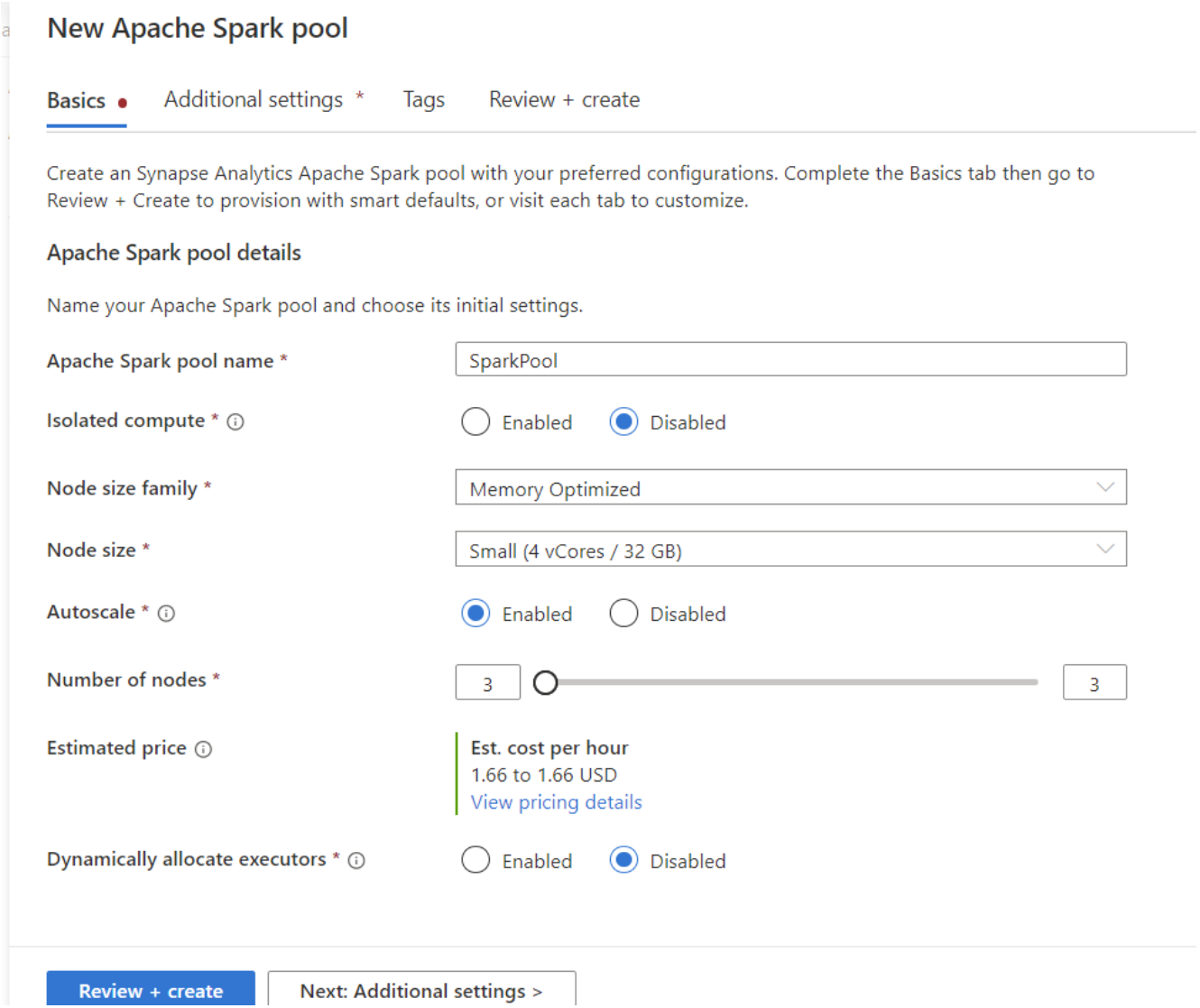
Data Ingestion with Apache Spark

- Thinking about the recurring task of data ingestion and a long-term solution, you may decided to perform data ingestion with an Apache Spark Notebook.
- First, upload the parquet file `parquet1.parquet` to your container
- Go to Synapse workspace and then click on `Apache spark pools` option under the `Manage` menu



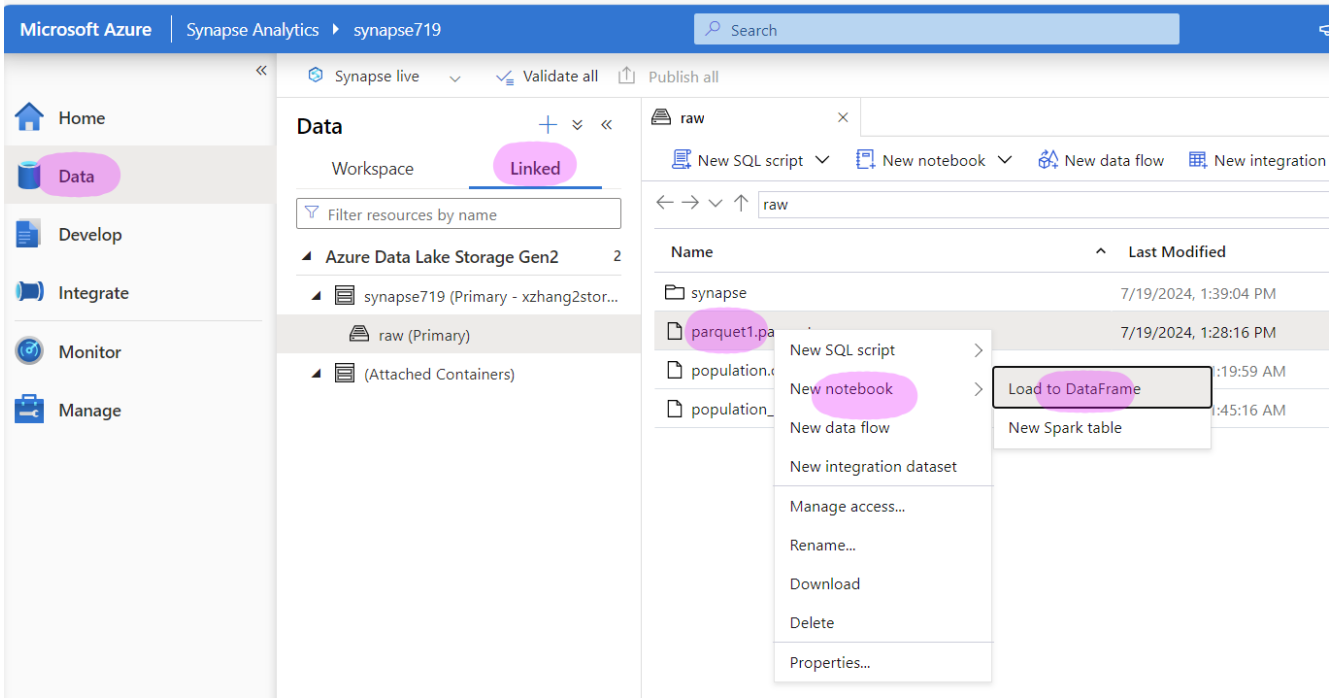
Data Ingestion with Apache Spark

- We need to create a new Apache Spark Pool



Data Ingestion with Apache Spark

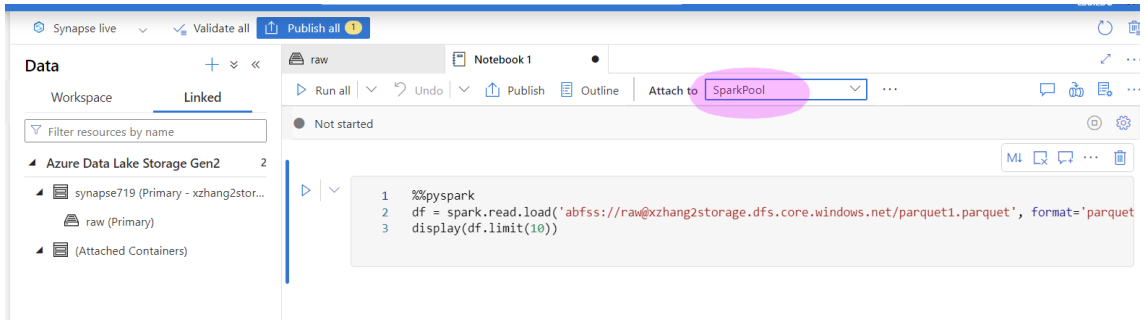
- After creating the Apache spark pool, Click on `Linked` services again in synapse studio.
  - Go to the parquet file, right click on it, click on the `New Notebook`, and select the `Load to DataFrame` option.



Data Ingestion with Apache Spark

- Your code will be like this
  - Run the code and check the result

```
%%pyspark
df = spark.read.load('abfss://raw@xzhang2storage.dfs.core.
              windows.net/parquet1.parquet', format='parquet')
display(df.limit(10))
```



Perform Data Transformation with DataFrames in Apache Spark Pools

- First, upload the JSON file `demo.json` to your container
- After loading the dataset, open synapse studio and follow the same steps that you did for the parquet file
  - Do not forget to attach the Apache spark pool that you have created.

```
%%pyspark
df = spark.read.load('abfss://raw@xzhang2storage.dfs.core.
                     windows.net/demo.json', format='json')
display(df.limit(10))
```

- Got some error message
  - Run the following, you will see `corrupt_record: string (nullable = true)`

```
df.printSchema()
```

Perform Data Transformation with DataFrames in Apache Spark Pools

- That is due to multiple lines of the JSON file. Update the code

```python
location = 'abfss://raw@xzhang2storage.dfs.core.windows.net/demo.json'

# Read the JSON file with the multiLine option set to True
df = spark.read.option("multiline", "true").json(location)

df.printSchema()
display(df.limit(10))
```

```python
location = 'abfss://raw@xzhang2storage.dfs.core.windows.net/demo.json'

# Read the JSON file with the multiLine option set to True
df = spark.read.option("multiline", "true").json(location)

df.printSchema()
display(df.limit(10))
```

[41] ✓ 3 sec - Command executed in 2 sec 820 ms by xzhang2 on 3:48:18 PM, 7/19/24

Job execution Succeeded **Spark** 2 executors 8 cores

```
root
 |-- CourseID: string (nullable = true)
 |-- CourseName: string (nullable = true)
 |-- UserId: string (nullable = true)
 |-- _attachments: string (nullable = true)
 |-- _etag: string (nullable = true)
 |-- _rid: string (nullable = true)
 |-- _self: string (nullable = true)
 |-- _ts: long (nullable = true)
 |-- id: string (nullable = true)
 |-- isBestCourse: boolean (nullable = true)
 |-- isPreferredCourse: boolean (nullable = true)
```

View Table Chart Export results

CourseID	CourseName	UserId	_attachments
DP203	Azure Data Engineering	1	attachments/
AZ-400	Azure Devops Expert	2	attachments/

Perform Data Transformation with DataFrames in Apache Spark Pools

- We can also create a view and run SQL queries on it.

```
df.createOrReplaceTempView("demo_views")
```

```
%%sql
SELECT * FROM demo_views
```

```
1  df.createOrReplaceTempView("demo_views")
```

[39] ✓ <1 sec - Command executed in 166 ms by xzhang2 on 3:47:28 PM, 7/19/24

```
1  %%sql
2  SELECT * FROM demo_views LIMIT 10
```

[40] ✓ 1 sec - Command executed in 510 ms by xzhang2 on 3:47:39 PM, 7/19/24

Job execution Succeeded **Spark** 2 executors 8 cores

View Table Chart Export results

CourseID	CourseName	UserId	_attachments
DP203	Azure Data Engineering	1	attachments/
AZ-400	Azure Devops Expert	2	attachments/

License