

# 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

- Create a Data Lake Storage account if you do not have one (Hierarchical namespace 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

Home > Microsoft Azure Synapse Analytics-20240719111558 | Overview > NetworkWatcherRG >

**synapse719**  
Synapse workspace

Search < > + New dedicated SQL pool + New Apache Spark pool + New Data Explorer pool (preview) Refresh Reset SQL admin password Delete


**Overview**


- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Settings
- Analytics pools
- Security
- Monitoring
- Automation
- Help

**Essentials**

Resource group (move)	: <a href="#">NetworkWatcherRG</a>	Networking	: <a href="#">Show firewall settings</a>
Status	: Succeeded	Primary ADLS Gen2 acco...	: <a href="#">https://xizhang2storage.dfs.core.windows.net</a>
Location	: East US	Primary ADLS Gen2 file s...	: raw
Subscription (move)	: <a href="#">Azure for Students</a>	SQL admin username	: sqladminuser
Subscription ID	: b5db3bdc-dae9-4e53-888f-8241b09dc786	SQL Microsoft Entra admin	: <a href="#">xizhang2@psu.edu</a>
Managed virtual network	: No	Dedicated SQL endpoint	: <a href="#">synapse719.sql.azuresynapse.net</a>
Managed Identity object ...	: 4d1506b3-4f32-4945-accf-49db4b7e68c3	Serverless SQL endpoint	: <a href="#">synapse719-ondemand.sql.azuresynapse.net</a>
Workspace web URL	: <a href="#">https://web.azuresynapse.net/workspaces/%7b5db3bdc-dae9-4e53-888f-8241b09dc7...</a>	Development endpoint	: <a href="#">https://synapse719.dev.azuresynapse.net</a>
Tags (edit)	: <a href="#">Add tags</a>		

**Getting started**

**Open Synapse Studio**  
Start building your fully-integrated analytics solution and unlock new insights.  
[Open](#)

**Read documentation**  
Learn how to be productive quickly. Explore concepts, tutorials, and samples.  
[Learn more](#)

# Prerequisite

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

[Home](#) > [Microsoft.Azure.SynapseAnalytics-20240719111558](#) | Overview > [NetworkWatcherRG](#) > [synapse719](#) >

## New dedicated SQL pool ...

[\\* Basics](#) [\\* Additional settings](#) [Tags](#) [Review + create](#)

Create a dedicated SQL 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. [Learn more](#)

### Dedicated SQL pool details

Name your dedicated SQL pool and choose its initial settings.

Dedicated SQL pool name \*

D\_SQLPool ✓

Performance level ⓘ

DW100c

DW100c

Estimated price ⓘ

Est. Cost Per Hour  
1.51 USD  
[View pricing details](#)

- 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

The screenshot shows the Microsoft Azure Synapse Studio interface for a workspace named 'synapse719'. The top navigation bar includes the Microsoft Azure logo, a search bar, and the workspace name. Below the navigation bar, there's a sidebar with various options: Overview (selected), Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings, Analytics pools, Security, Monitoring, Automation, and Help. The main content area is divided into two sections: 'Essentials' and 'Getting started'. The 'Essentials' section displays key information about the workspace, including its resource group, status, location, subscription, and various endpoints. The 'Getting started' section provides links to 'Open Synapse Studio' and 'Read documentation'. Below these sections, there's a table titled 'Analytics pools' which lists the available pools and their configurations.

Microsoft Azure

Home > Microsoft.Azure.Synapse.SqlPoolOnExistingWorkspace\_5beb193ada9b4 | Overview >

**synapse719**  
Synapse workspace

Search resources, services, and docs (G+I)

Search

+ New dedicated SQL pool + New Apache Spark pool + New Data Explorer pool (preview) Refresh Reset SQL admin password Delete

### Essentials

Resource group (mspub)	: <a href="#">NetworkWebster55</a>	Networking	: <a href="#">Show firewall settings</a>
Status	: Succeeded	Primary ADLS Gen2 acco...	: <a href="#">https://xzhang2storage.dfs.core.windows.net</a>
Location	: East US	Primary ADLS Gen2 file s...	: raw
Subscription (mspub)	: <a href="#">Azure for Students</a>	SQL admin username	: sqladminuser
Subscription ID	: b5db1bdc-dbe9-4e53-888f-8241a09dc706	SQL Microsoft Entra admin	: <a href="#">xzhang2@asu.edu</a>
Managed virtual network	: No	Dedicated SQL endpoint	: <a href="#">synapse719.sql.azure.synapse.net</a>
Managed Identity object ...	: 4d8506b3-4f52-4945-acc1-49db607e68c3	Serverless SQL endpoint	: <a href="#">synapse719-ondemand.sql.azure.synapse.net</a>
Workspace web URL	: <a href="#">https://web.azure.synapse.net/workspace=%7B%22subscriptions%22%3A%22b5db1bdc-dbe9-4e53-888f-8241a09dc706%7D</a>	Development endpoint	: <a href="#">https://synapse719.dev.azure.synapse.net</a>
Tags (add)	: <a href="#">Add tags</a>		

### Getting started

Open Synapse Studio  
Start building your fully integrated analytics solution and unlock new insights.

[Open](#)

Read documentation  
Learn how to be productive quickly. Explore concepts, tutorials, and samples.

[Learn more](#)

### Analytics pools

Search to filter items...

Name	Type	Size
<b>SQL pools</b>		
Built-in	Serverless	Auto
D_SqlPool	Dedicated	DW100c
<b>Apache Spark pools</b>		
No pools provisioned		

## Prerequisite

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

Microsoft Azure

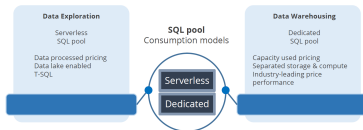
Synapse



Analytics pools

# 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.

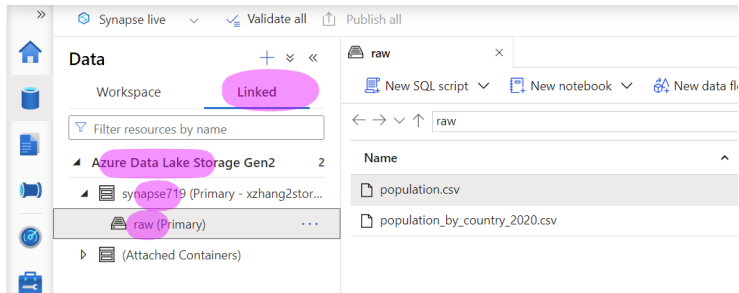


- A **synapse notebook** is an interface used to create a file consisting of live code and text by validating ideas and fetch insights from the data.



# 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.

The screenshot shows the Synapse Studio interface for previewing a CSV file. On the left, a sidebar shows a file explorer with a 'raw' folder and a file named 'population\_by\_country\_2020.csv'. The main area displays the file's details and a preview of its contents.

**population\_by\_country\_2020.csv**

**Path** `https://xzhang2storage.dfs.core.windows.net/raw/population_by_country_2020.csv`

**Modified** 7/19/2024, 11:45:16 AM

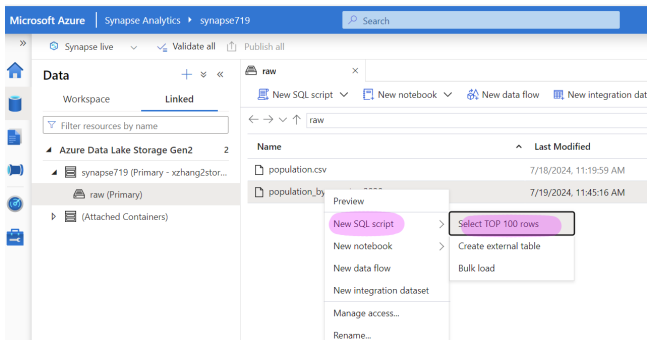
**With column header** ☒ On

COUNTRY	POPULATION...	YEARLY_CHA...	NET_CHA...
China	1440297825	0.39%	5540090
India	1382345085	0.99%	13586631
United States	331341050	0.59%	1937734
Indonesia	274021604	1.07%	2898047
Pakistan	221612785	2.00%	4327022
Brazil	212821986	0.72%	1509890
Nigeria	206984347	2.58%	5175990
Bangladesh	164972348	1.01%	1643222

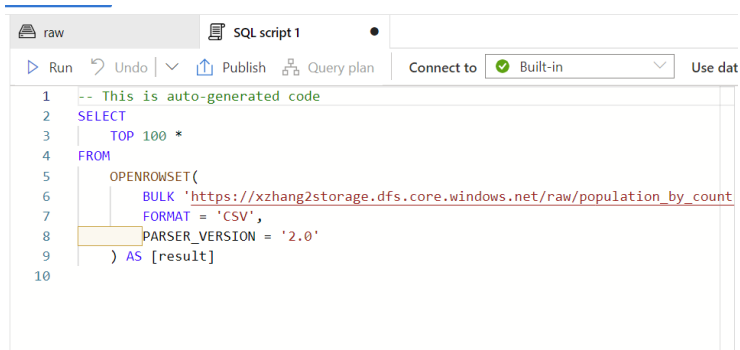
At the bottom of the preview window is an 'OK' button.

# 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



The screenshot shows the Synapse Studio interface. At the top, there are tabs for 'raw' and 'SQL script 1'. Below the tabs is a toolbar with buttons for 'Run', 'Undo', 'Publish', and 'Query plan'. To the right of the toolbar is a 'Connect to' dropdown menu set to 'Built-in' and a 'Use dat' button. The main area is a SQL editor with the following code:

```
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          PARSE_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.



raw



Run



Undo

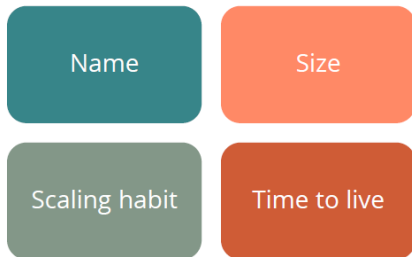
1

--

This is

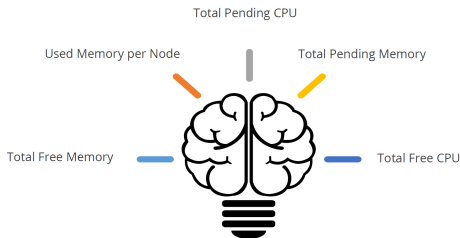
# Apache Spark in Azure Synapse Analytics

- **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:



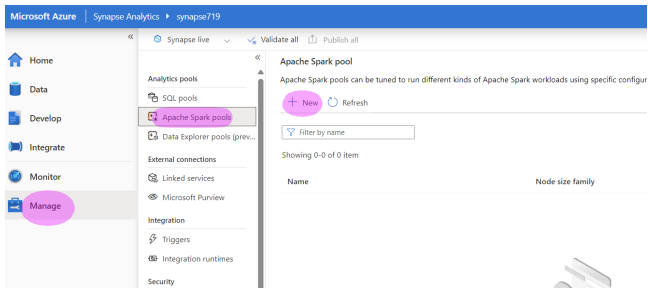
# Apache Spark in Azure Synapse Analytics

- 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 decide 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

## New Apache Spark pool

Basics • Additional settings \* Tags Review + create

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 \*

SparkPool

Isolated compute \* ⓘ

☐ Enabled ☒ Disabled

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  
1.66 to 1.66 USD  
[View pricing details](#)

Dynamically allocate executors \* ⓘ

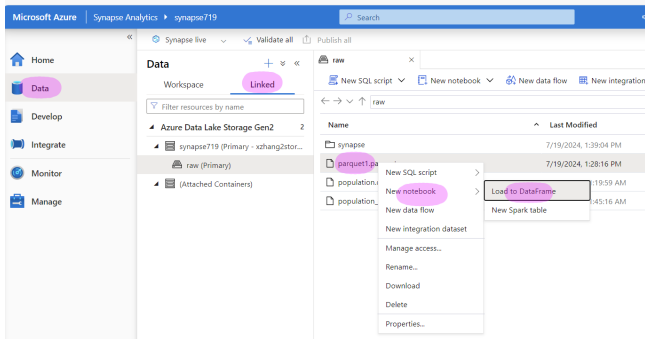
☐ Enabled ☒ Disabled

Review + create

Next: Additional settings >

# Data Ingestion with Apache Spark

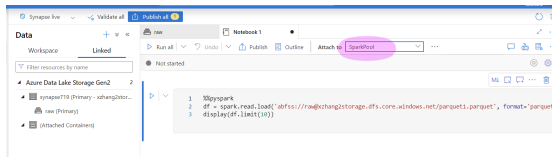
- After creating the Apache spark pool, Click on Linked services again in synapse studio.
  - ▶ Go to the parquet file, right click on it, click on the New Notebook, and select the Load to DataFrame option.



# 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

```
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))
```

The screenshot shows a Databricks notebook cell with the following code:

```
1 location = "abfss://raw@xzhang2storage.dfs.core.windows.net/demo.json"
2
3 # Read the JSON file with the multiline option set to True
4 df = spark.read.option("multiline", "true").json(location)
5
6 df.printSchema()
7 display(df.limit(10))
```

Below the code, the execution status is shown as "Job execution Succeeded" with "Spark 2 executors 8 cores".

The schema of the DataFrame is displayed as follows:

```
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)
```

At the bottom, there is a table view of the data with columns: CourseID, CourseName, UserID, and \_attachments. The table shows two rows of data.

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   [Export results](#) ▼

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

# License



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#).