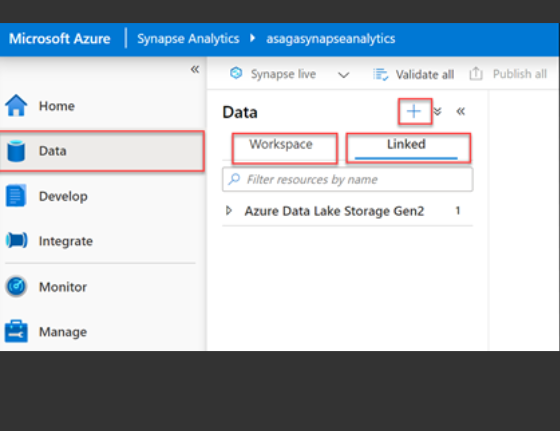
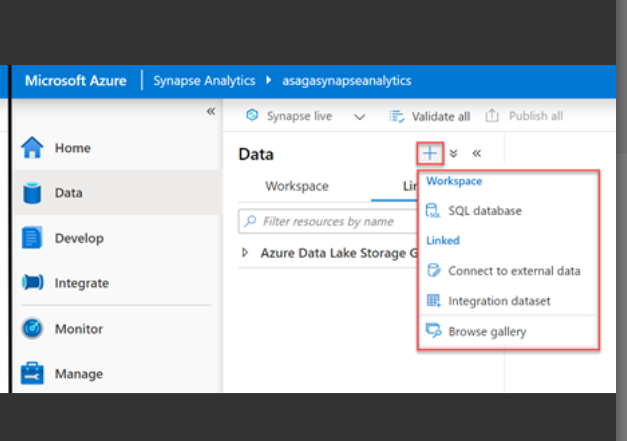
https://atos365-my.sharepoint.com/:f:/r/personal/muchkund\_tiwari\_atos\_net/Documents/Documents/Udemy/nyc\_green\_taxi\_data/raw?csf=1&web=1&e=5AJLns

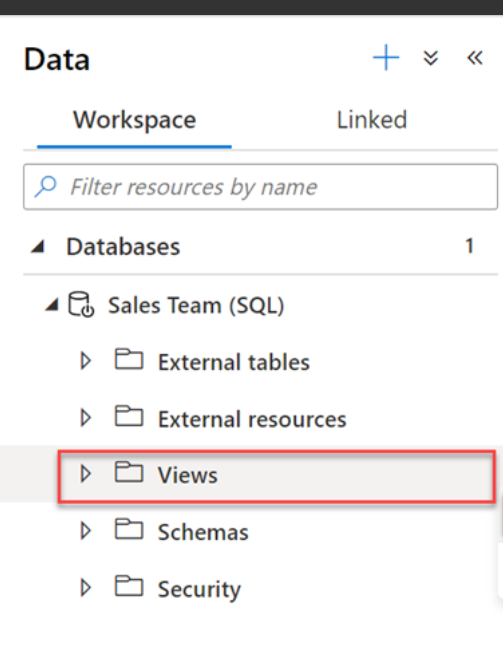
Datahub in azure synapse analytics

There are two tabs in the Data Hub blade, **Workspace** and **Linked**. The **Workspace** tab is where you will find your workspace databases for both SQL and Spark.



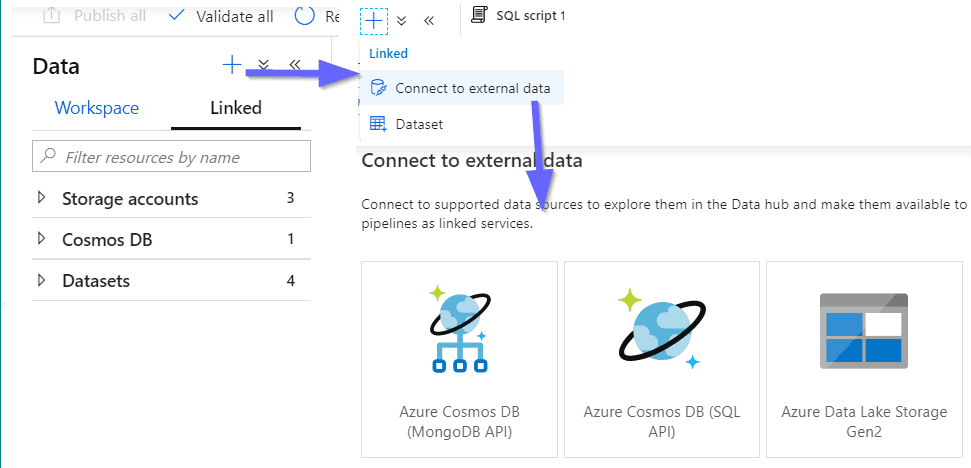


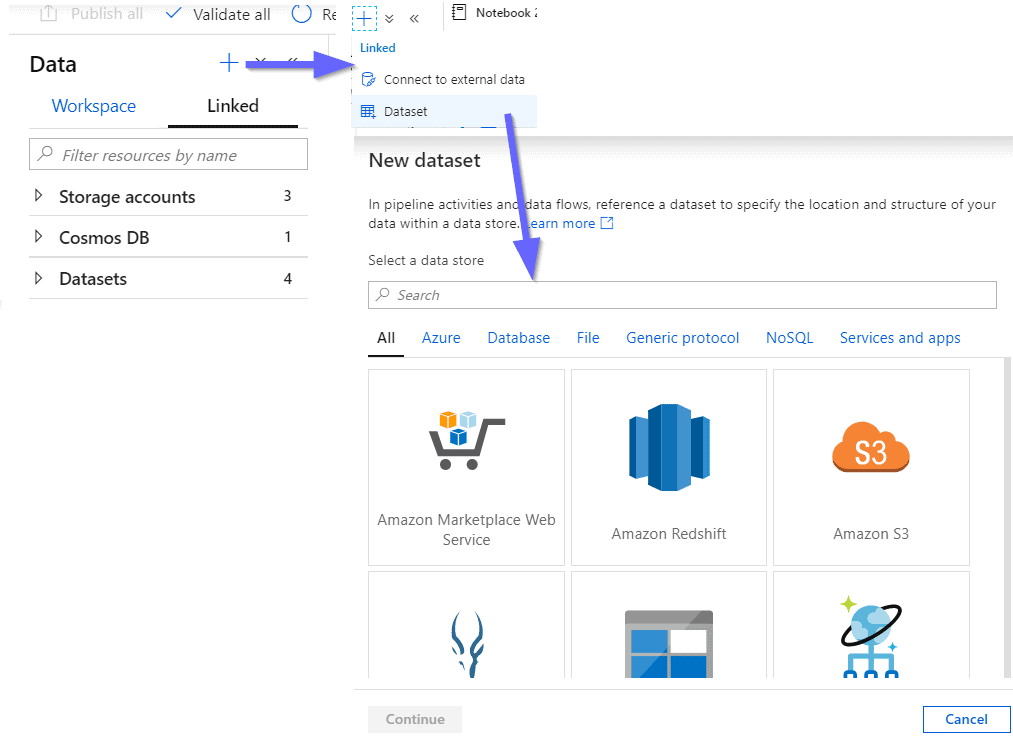
After click on plus we have multiple option from them we choose the SQL database, and it will show two option serverless and dedicated pool

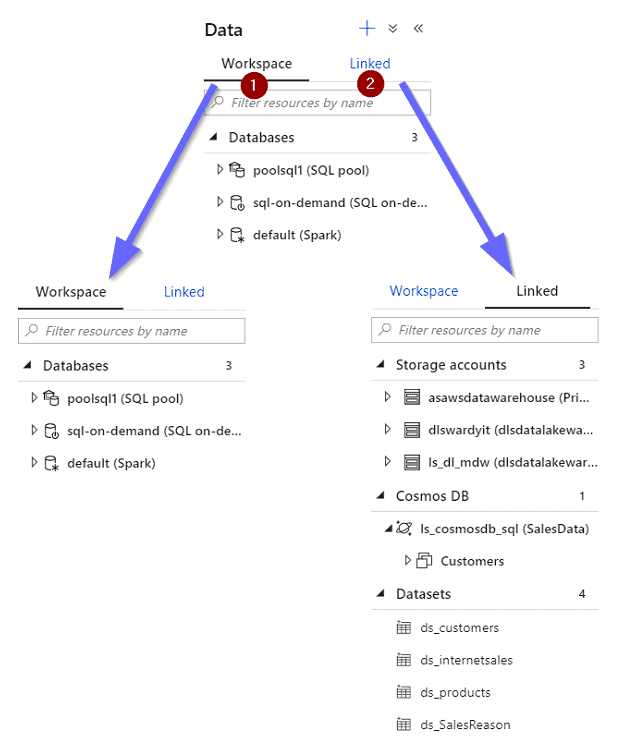


Here the Database\_name will be the Database under that there will be sales database and now you can create External Table, Views etc. underneath that database.

If you go to linked section

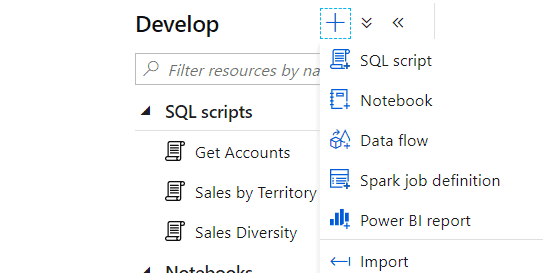


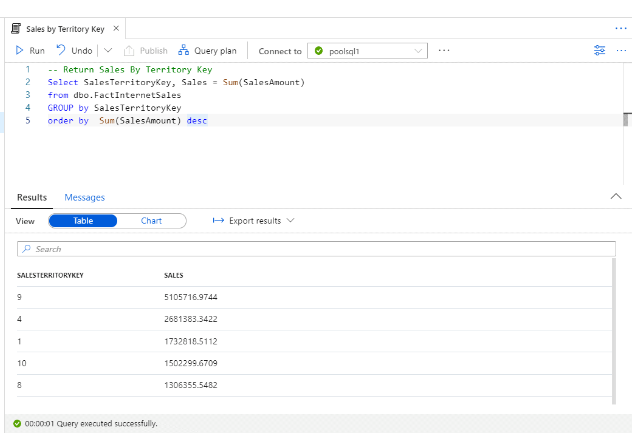




[Azure Synapse Analytics Data Hub - David Alzamendi](https://davidalzamendi.com/azure-synapse-analytics-data-hub/)

Develop in azure synapse analytics

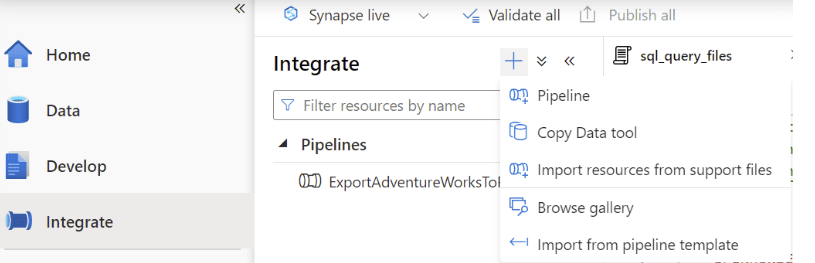
SQL Script: -



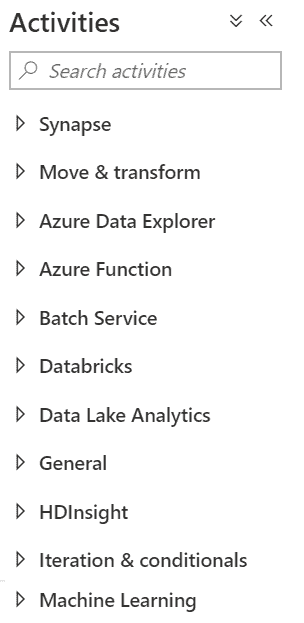
Here we connect to a SQL Pool and here it will be poolsql1 and it is in Table format

[Azure Synapse Analytics Develop Hub - David Alzamendi](https://davidalzamendi.com/azure-synapse-analytics-develop-hub/#:~:text=With%20the%20Develop%20Hub%20in%20Azure%20Synapse%20Analytics,Flows%20%28Azure%20Data%20Factory%29%206%20Spark%20Job%20Definitions) for more type we can visit this page

Integrate in azure synapse analytics

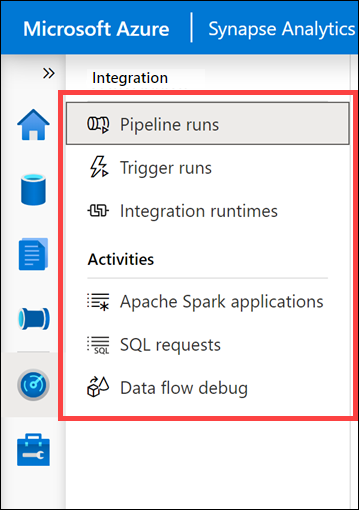


After click on pipeline, it will show like this



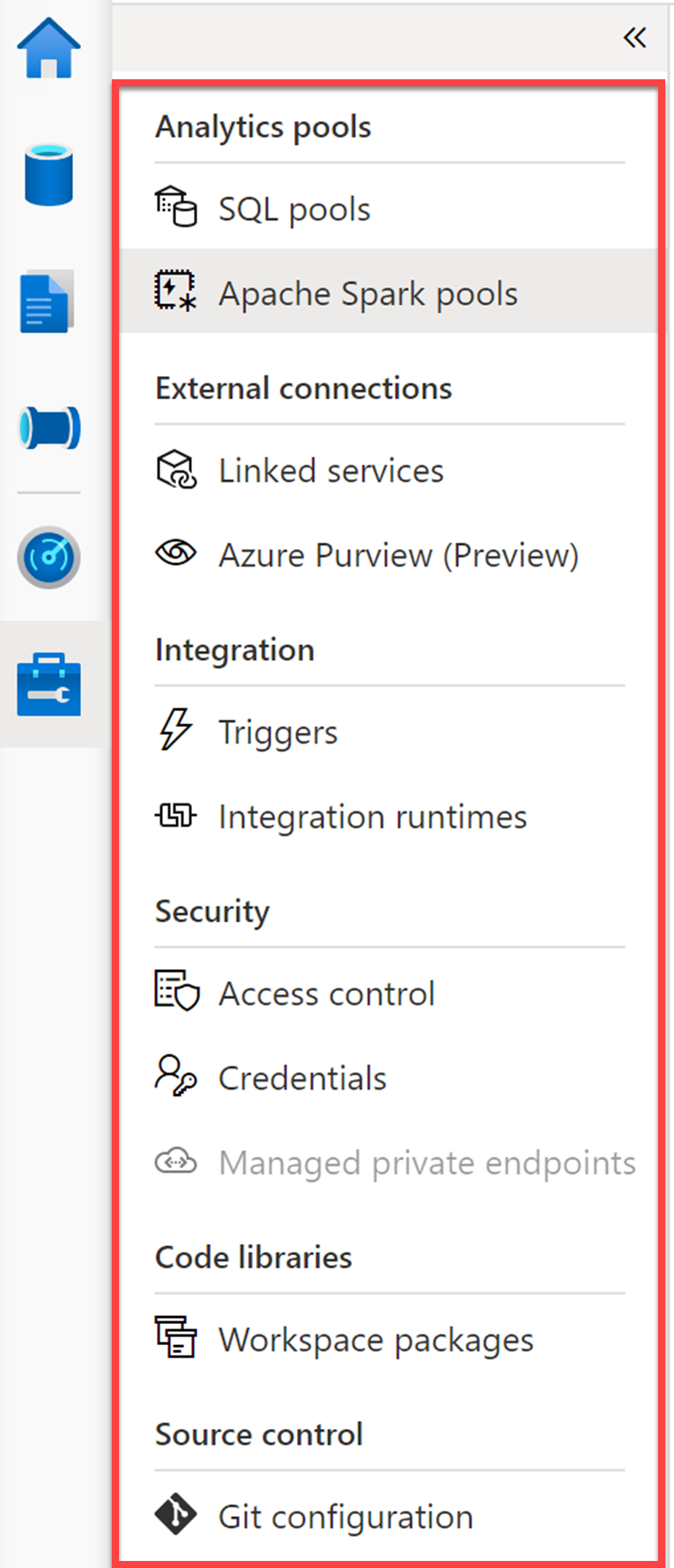
[Azure Synapse Analytics Integrate Hub - David Alzamendi](https://davidalzamendi.com/synapse-analytics-integrate-hub/)

Monitor in azure synapse analytics



[Explore the Monitor hub - Training | Microsoft Learn](https://learn.microsoft.com/en-us/training/modules/explore-azure-synapse-studio/10-explore-monitor-hub)

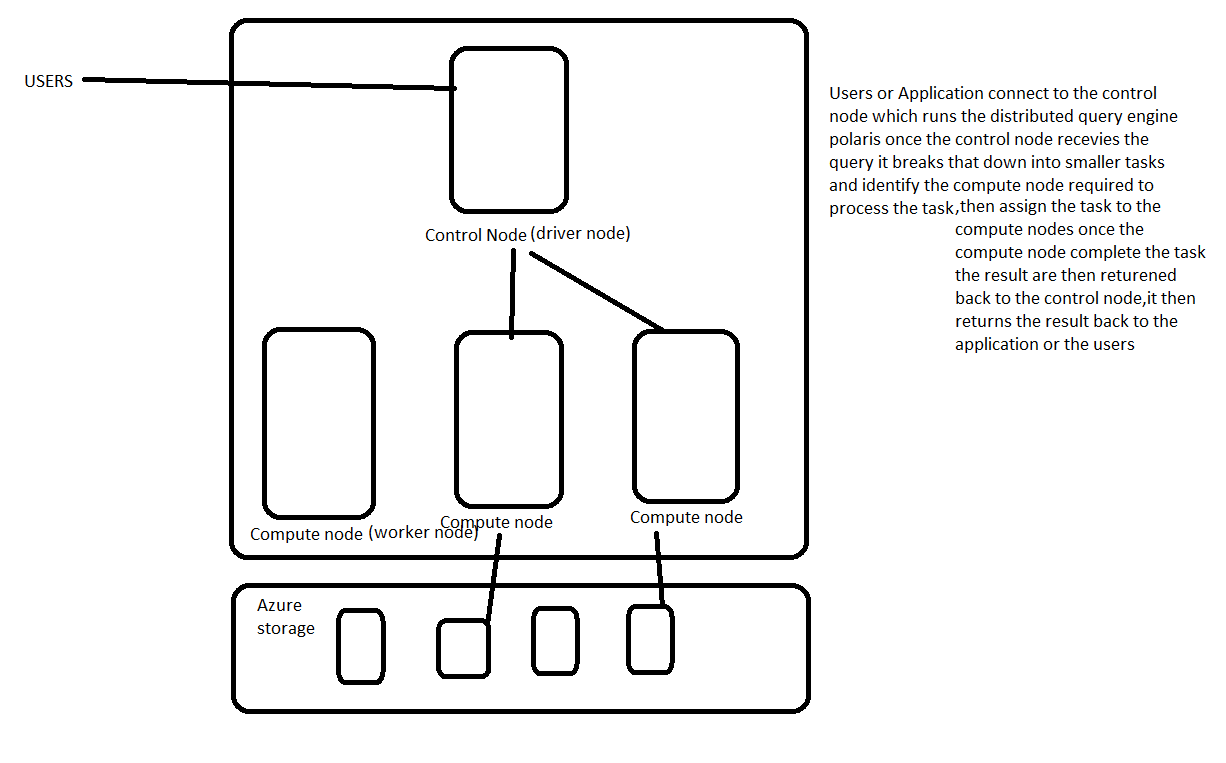
Manage in azure synapse analytics



Serverless SQL pool- Architecture

SQL Pool is a feature in Azure Synapse Analytics, which is a cloud-based analytics service. SQL Pool is a fully managed, enterprise-level data warehousing solution that enables organizations to store and manage large amounts of structured and semi-structured data for analytics purposes.

Serverless SQL pool is a serverless distributed query engine that you can use to query data over your data lake using T-SQL

Key feature = serverless (Hence no infrastructure to provision our cluster to administer)

Use distributed query engine called Polaris

Robust =The distributed query engine has built in query execution for **fault tolerance (**Fault tolerance refers to the ability of a system (computer, network, cloud cluster, etc.) to continue operating without interruption when one or more of its components fail.) and it provides high reliability and query success rates, even for long running queries involving large data sets it also add and remove the node as required by the query.

Query using T-SQL, pay-per-query pricing model (serverless sql POOL MAY USE ONE OR MORE COMPUTE NODES FOR EXECUTING OUR QUERIES, AND THEY COULD BE ANY SIZE as a consumer we don't pay for the resources directly so there isn’t any upfront cost the billing is based on the **data being processed by the queries we have run**) tho who query Kitna amount of data processed kar raha ha ispar depend ha

Not store any data, Synapse link use

Serverless Sql pool – Supported Data Sources

It supports both azure blob storage and azure data lake storage

Azure Storage account= Delimited-CSV, TSV, etc

JSON, Parquet, Delta Lake

Cosmos DB= read cosmos DB data via synapse link (SQL Api, Mongo db. Api)

You can also read from Dataverse via synapse link

SQL Server 2022(preview)

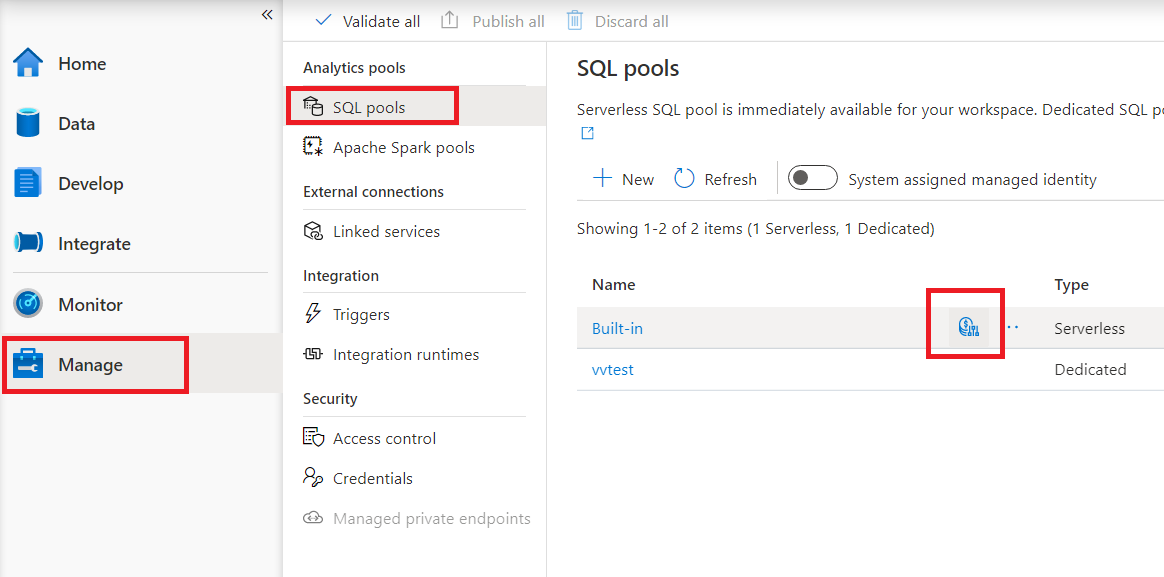
Cost Control

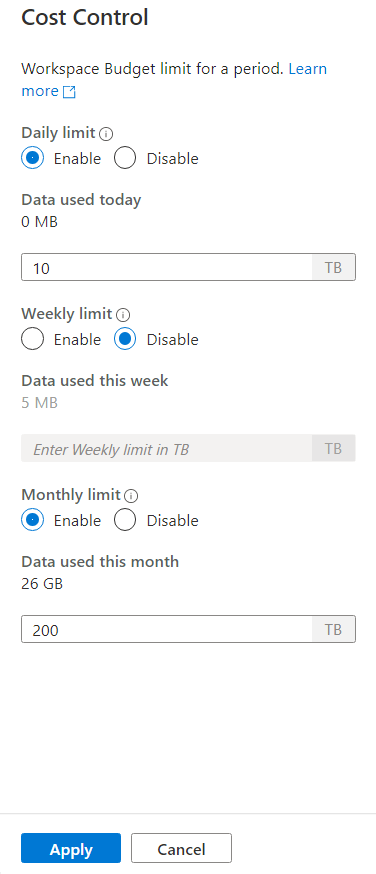
Billed for Data Processes =1) The amount of data read from the storage the data read from the storage will be dependent on the query that you’re writing and the format of the files ex:- if you partitioned your data and the query required only certain partitions to be read serverless SQL will only read those partitions and you are only charged for that amount of data. Instead, if the data is not partitioned it may have to read all the data every time 2) some queries will required the data to be transferred amongst compute nodes as part of the query execution this data will also be included for billing 3) we can use external table statements to write out the data to storage using serverless

Data processed doesn’t include server level metadata such as logins and also executions of DDL statement such as create database, create schema, etc.

Data processed rounded to nearest MB

minimum of 10 mb be charged per query

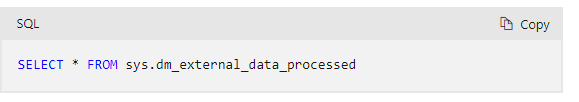




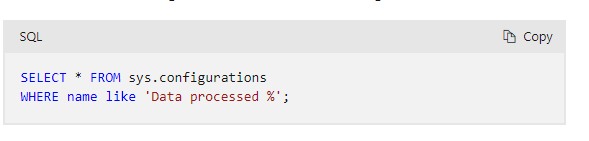
It shows how much data has been processed by this workspace today, this week and this month and it’s gives you the ability to put control or put limit on the amount of data the workspace is allowed to process

Now we perform the cost control through Interface let’s do now through T-SQL

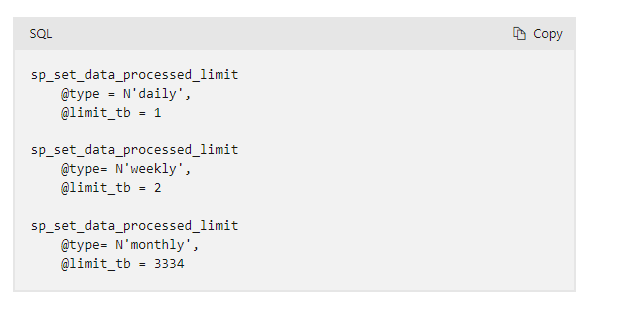
Go to the develop in that create SQL script



To see how much data we processed in daily, weekly and monthly we use above query



To see the current setting for the daily, weekly, monthly limits we use above query



To set the limits we use above query

Query CSV file

OPENROWSET Function

The easiest way to see to the content of your CSV file is to provide file URL to OPENROWSET function, specify csv FORMAT, and 2.0 PARSER\_VERSION.

The OPENROWSET function can be referenced in the FROM clause of a query as if it were a table name

Syntax: -

SELECT \*

FROM OPENROWSET (BULK ‘blob file path’,

FORMAT = ['CSV’, ‘PARQUET’, ‘DELTA’]

) AS [FILE]

Example: -



Use **PARSER\_VERSION 2.0** to query CSV files You can use a performance-optimized parser when you query CSV files.

[How to use OPENROWSET in serverless SQL pool - Azure Synapse Analytics | Microsoft Learn](https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset)

Query a TSV File

Example how to query TSV file in sql script (it is not separated by comma, it is separated by tap)

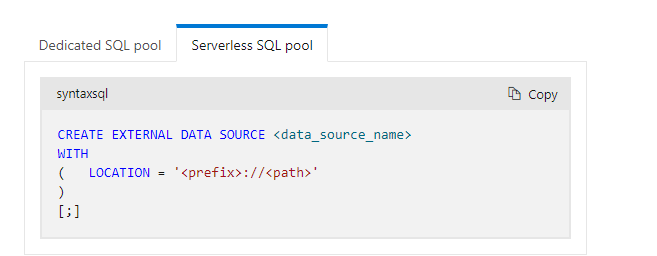


Debugging & Identifying error

If you are getting an error so it is difficult to understand that error and also unable to identify the exact problem or pinpoint as to the exact location of where the problem is and it is very difficult, so this is the limitation of PARSER\_VERSION = ‘2.0’ so the solution is PARSER\_VERSION = ‘1.0’

External Data Source

The URL with BULK is quite tedious to write if you do it by hand and also, it’s not an easy one to Rember so you don’t want your developer to be writing this every time, they want to query the files also the URL may have to be changed every time we move from one end to the other example:- if you have separate storage accounts for dev, test, production environment you will have to make the changes to all your queries whenever you want to use them in a new environment which is not idle that’s where external data source come into play they help us hide the URL and give it a meaningful name



Example: -

CREATE EXTERNAL DATA SOURCE nyc\_taxi\_data\_raw {by using this data source I can access all the data that belong to our project}

WITH (

LOCATION =‘ abfss ://nyc-taxi-data @syncapsecoursedl.dfs.core.windows.net/raw’

) {we provide a location to which our data source pointing, when you create this data source that will directly put you on the raw folder so you don’t have to specify the raw folder when you select the data}

SELECT \*

FROM OPENROWSET (

BULK ‘taxi\_zone.csv’, {I can access data through this amount of URL}

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,{you must provide data source name again, to access data from the data source you would specify the data source which is pointing up to raw }

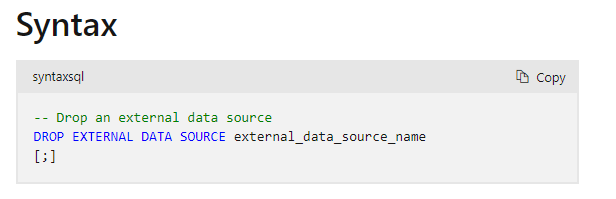
FORMAT = ‘CSV’

PARSER\_VERSION = ‘2.0’

)

[CREATE EXTERNAL DATA SOURCE (Transact-SQL) - SQL Server | Microsoft Learn](https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=serverless)

Delete External Data source



Data Type

To see the data type of my column we have to write

EXEC sp\_describe\_first\_result\_set N ‘pass your sql query and we need to double quote them

* Use WITH caluse to provide explicit data types

[DROP EXTERNAL DATA SOURCE (Transact-SQL) - SQL Server | Microsoft Learn](https://learn.microsoft.com/en-us/sql/t-sql/statements/drop-external-data-source-transact-sql?view=sql-server-ver16)

Some problem we face

Suppose we have a CSV file in that there is two column heading but we have three data

Vendor\_id, Vendor\_name (Two heading we have)

1, Creative Mobile Technologies, LLC (But the data is Three)

2, Verifone Inc

Solution: - Use escape character (\) so that common will not consider. (\\) because slash should also be escape or put all the data in single string except number

Example: -

first way = 1, Creative Mobile Technologies\, LLC (ESCAPECHAR = ‘\\’) what happen in that every string has been coated

Second way = 1, “Creative Mobile Technologies, LLC” (FIELDQUOTE = ‘ “ ’) by default field terminator

Query JSON

Line Delimited JSON = In line delimited json file where json document are separated by a new line character this kind of json file generally found in telemetry-based systems



Standard JSON = In this type of json document, multiple json document are stored as on single array

Example : -

[

{“rate\_code\_id”:1, “rate\_code”: “standard rate”},

{“rate\_code\_id”:1, “rate\_code”: “JFK”}

]

Classic JSON = This are very similar to the standard json in that many json document stored as one single array the difference here is that individual json document are nicely formatted into multiple line



As we saw previously, openrowset function doesn’t have a specific parser for JSON document, but we can use a combination of CSV parser and a couple of function to query json file using serverless pool as you know in a json document, each element and value are separated by a comma so if we leave the default field terminator of comma, it’ll split the json document into many little pieces, which is not we want also, the json document can have quotes in the data, so we need to override the field terminator and the field quote with a character that’s generally not seen in the data

Line-delimited JSON

FIELDTERMINATOR – 0x0b

FIELDQUOTE – 0X0b

Standard JSON

FIELDTERMINATOR – 0x0b

FIELDQUOTE – 0X0b

ROWTERMINATOR – 0X0b (vertical tab)

Now that the openrowset function output each record as a valid json document, we can feed that into a json function to pass the json and extract the element into columns. We have two function 1) **JSON\_VALUE** and we use this to extract the element names and values into column names and values it has some limitation that it can't handle arrays and also, it’s inefficient when dealing with large JSON document with multiple prosperities for this we can use function called 2) **OPENJSON** which will explode array for us.

Query Payment Type

select CAST(JSON\_VALUE (jsonDOC, '$.patment\_type') AS SMALLINT) payment\_type, (we want to pull out the property payment type from this json which is jsonDOC )

CAST (JSON\_VALUE (joinDOC, '$.payment\_type\_desc') AS VARCHAR (15)) payment\_type\_desc

from openrowset(

bulk 'payment\_type\_json',

data\_source = 'nyc\_taxi\_data\_raw',

format = 'csv',

PARSER\_VERSION = ‘1.0’, (if you’re using parser\_version 1.0 then you have to specify the column name)

fieldterminator ='0x0b',

fieldquote = '0x0b'

)

WITH

(

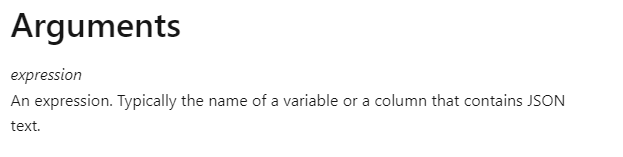
jsonDoc NVARCHAR(MAX) (we are going to read all the data from JSON document into one field which is jsonDoc in this case)

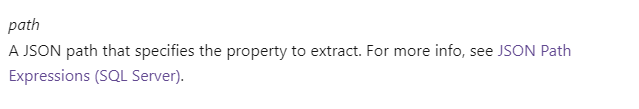
) AS payment type;

($) represent this

(CAST) it is function which is used to specify the data type







[JSON\_VALUE (Transact-SQL) - SQL Server | Microsoft Learn](https://learn.microsoft.com/en-us/sql/t-sql/functions/json-value-transact-sql?view=sql-server-ver16)

OPENJSON is a table valued function which take in a json document and turns its object and properties into rows and columns, and also with the open json function, you can specify data types instead of using cast in your select statement and also, it’ll explode array when you have them within your data or within your Json document and also with in the width class you can explicitly list the columns and you can rename the columns if required and you can specify the data as well.

SELECT payment\_type, payment\_type\_desc

from openrowset(

bulk 'payment\_type\_json',

data\_source = 'nyc\_taxi\_data\_raw',

format = 'csv',

PARSER\_VERSION = ‘1.0’, (if you’re using parser\_version 1.0 then you have to specify the column name)

fieldterminator ='0x0b',

fieldquote = '0x0b'

)

WITH

(

jsonDoc NVARCHAR(MAX)

) AS payment type;

CROSS APPLY OPENJSON (jsonDOC) = what we done till now is we take json document that comes out from openrowset and push that in openjson function so the openjson function will pass the json document and get the properties and values so we going to get some key: values and you can turn that key: values into column name

(CROSS APPLY) will work as join

WITH

(

Payment\_type SMALLINT,

Payment\_type\_desc VARCHAR (20)

);

Query Standard JSON

In the standard json file in which multiple JSON document are grouped into JSON array in the standard json each of the lines are not really valid JSON since it will start from comma and then you’ve got a JSON structure

SELECT rate\_code\_id, rate\_code

from openrowset(

bulk 'rate\_code.json,

data\_source = 'nyc\_taxi\_data\_raw',

format = 'csv',

PARSER\_VERSION = ‘1.0’, (if you’re using parser\_version 1.0 then you have to specify the column name)

FIELDTERMINATOR – 0x0b

FIELDQUOTE – 0X0b

ROWTERMINATOR – 0X0b

)

WITH

(

jsonDoc NVARCHAR(MAX)

) AS rate\_code

CROSS APPLY OPENJSON (jsonDOC)

WITH

(

rate\_code\_id TINYINT,

rate\_code VARCHAR (20)

);

Through the JSON\_VALUE we can pass the index also.

Example:-

CAST (JSON\_VALUE(jsonDoc , ‘$.payment\_type\_desc[0].value’) AS VARCHAR (15)) payment\_type\_desc { in this the index is 0 and the object is value} but it is not valid method because the main reason is when you have arrays in JSON files, they’re not necessarily having a limited number of occurrences they could be unbounded so it’ll be impossible for us to write query to turns those into columns

Select the Data from Folder

CREATE EXTERNAL DATA SOURCE nyc\_taxi\_data\_raw {by using this data source I can access all the data that belong to our project}

WITH (

LOCATION =‘ https://syncapsecoursedl.dfs.core.windows.net/nyc-taxi-data/raw/’

)

SELECT

TOP 100 \*

FROM

openrowset(

BULK 'trip\_data\_green\_csv/year=2020/month=01/’,

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,

FORMAT = ‘CSV',

PARSER\_VERSION = ‘2.0’

HEADER\_ROW = TRUE

) AS [result]

If you have multiple files in the folder

SELECT

TOP 100 \*

FROM

openrowset(

BULK 'trip\_data\_green\_csv/year=2020/month=01/\*’,

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,

FORMAT = ‘CSV',

PARSER\_VERSION = ‘2.0’

HEADER\_ROW = TRUE

) AS [result]

If you have multiple type of file but you want to select all the data from CSV file then

SELECT

TOP 100 \*

FROM

openrowset(

BULK 'trip\_data\_green\_csv/year=2020/month=01/\*.CSV’,

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,

FORMAT = ‘CSV',

PARSER\_VERSION = ‘2.0’

HEADER\_ROW = TRUE

) AS [result]

Actually, we want to select data from all the files and the year 2020 so that would be the file in the month 01,02,03 up to 12 but there are no data files under the year 2020 they are sub folder

Select data from subfolders

SELECT

TOP 100 \*

FROM

openrowset(

BULK 'trip\_data\_green\_csv/year=2020/\*\*’, (If you are going to get data from subfolder recursively, then we have to specify \*\*)

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,

FORMAT = ‘CSV',

PARSER\_VERSION = ‘2.0’

HEADER\_ROW = TRUE

) AS [result]

We want to get data from a couple of months, not every month

SELECT

TOP 100 \*

FROM

openrowset(

BULK ( 'trip\_data\_green\_csv/year=2020/month=01/\*.CSV’,

'trip\_data\_green\_csv/year=2020/month=02/\*.CSV’),

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,

FORMAT = ‘CSV',

PARSER\_VERSION = ‘2.0’

HEADER\_ROW = TRUE

) AS [result]

Use more than 1 wildcard character getting data from all the years and all the months

SELECT

TOP 100 \*

FROM

openrowset(

BULK 'trip\_data\_green\_csv/year=\*/month=\*/\*.CSV’,

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,

FORMAT = ‘CSV',

PARSER\_VERSION = ‘2.0’

HEADER\_ROW = TRUE

) AS [result]

How to get the file name alongside each of the record so that we can know which record came from which file so to get that serverless SQL pool gives files meta data function called filename ()

SELECT

TOP 100

result.filename() AS file\_name,

result. \*

FROM

openrowset(

BULK 'trip\_data\_green\_csv/year=\*/month=\*/\*.CSV’,

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,

FORMAT = ‘CSV',

PARSER\_VERSION = ‘2.0’

HEADER\_ROW = TRUE

) AS [result]

Limit the Data using filename()

SELECT

TOP 100

result.filename() AS file\_name,

result. \*

FROM

openrowset(

BULK 'trip\_data\_green\_csv/year=\*/month=\*/\*.CSV’,

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,

FORMAT = ‘CSV',

PARSER\_VERSION = ‘2.0’

HEADER\_ROW = TRUE

) AS [result]

WHERE result.filename() IN (‘green\_tripdata\_2020-01.csv’, ‘green\_tripdata\_2021-01.csv’)

GROUP BY result.filename(),

ORDER BY result.filename();

You get the path of the file through filepath()

SELECT

TOP 100

result.filename() AS file\_name,

result.filepath() AS file\_path,

result. \*

FROM

openrowset(

BULK 'trip\_data\_green\_csv/year=\*/month=\*/\*.CSV’,

DATA\_SOURCE = ‘nyc\_taxi\_data\_raw’,

FORMAT = ‘CSV',

PARSER\_VERSION = ‘2.0’

HEADER\_ROW = TRUE

) AS [result]

GROUP BY result.filepath(),

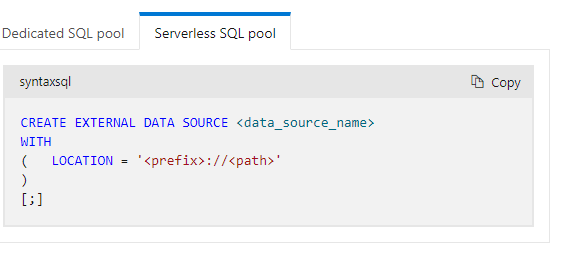
ORDER BY result.filepath();

Data Virtualization

[CREATE EXTERNAL DATA SOURCE (Transact-SQL) - SQL Server | Microsoft Learn](https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&tabs=dedicated)

External data source

In Azure Synapse Analytics, an external data source is a way to connect to data outside of your Synapse workspace, such as data stored in Azure Blob Storage, Azure Data Lake Storage, Azure SQL Database, or other data stores.



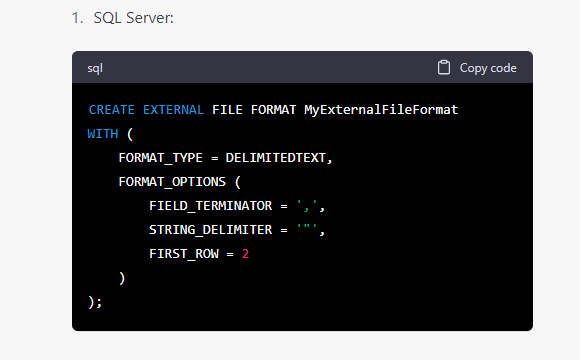
https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&tabs=dedicated

External File format

In Azure Synapse Analytics, an external file format is used to define the structure and properties of data stored in external files, such as CSV, JSON, Parquet, or ORC files. External file formats allow you to describe how the data is organized and formatted, such as the field and row delimiters, the encoding type, and the data types of each column.



Delimited text



[CREATE EXTERNAL TABLE (Transact-SQL) - SQL Server | Microsoft Learn](https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated)

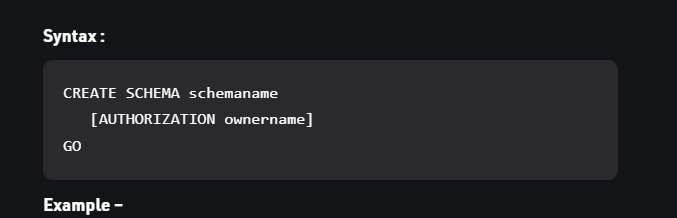
External Table

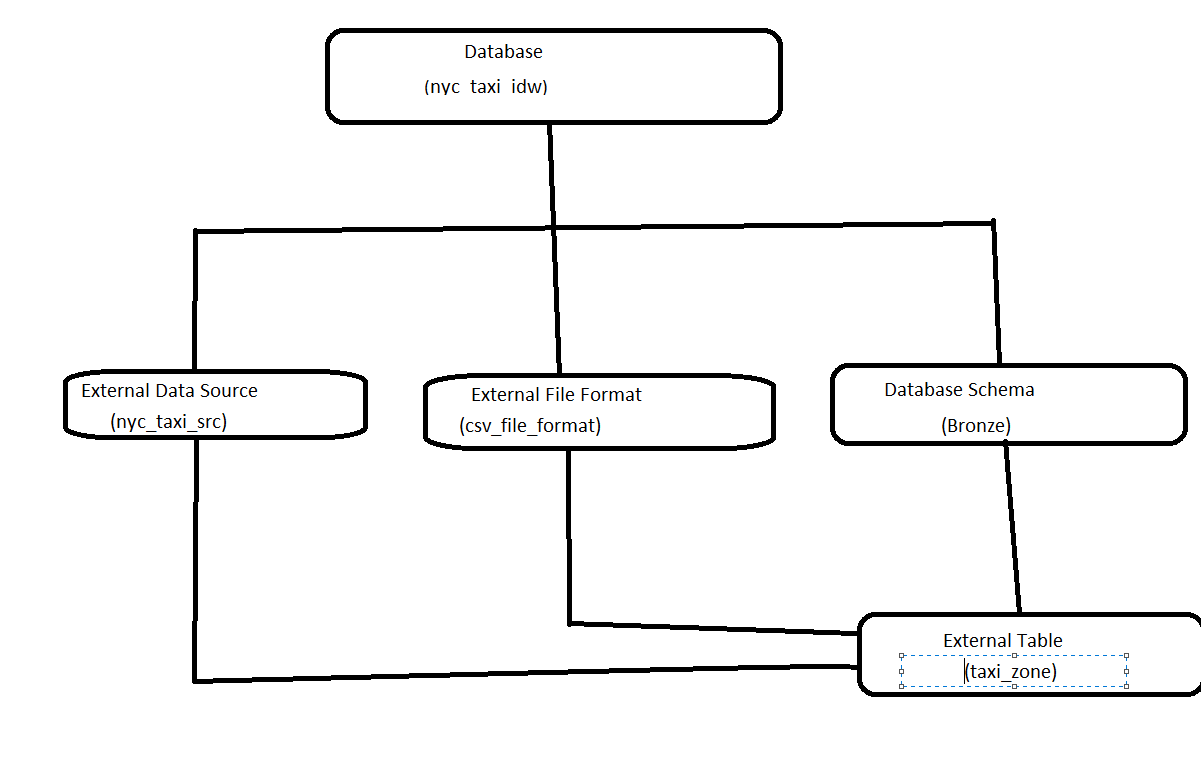




Create Schema

In Azure Synapse Analytics, a schema is a container for database objects such as tables, views, functions, and stored procedures. A schema is a way to organize database objects into logical groups, which can help with security, manageability, and performance.





Create external file format for parquet\_file\_format

IF NOT EXISTS (SELECT \* FROM sys.external\_file\_format WHERE name = ‘parquet\_file\_format’)

CREATE EXTERNAL FILE FORMAT parquet\_file\_format

WITH (

FORMAT\_TYPE = PARQUET,

DATA\_COMPRESSION = ‘org.apache.io.compress.snappyCodec’

);

File format is created now we must create External Table

IF OBJECT\_ID(‘bronze.trip\_data\_green\_parquet’) IS NOT NULL

DROP EXTERNAL TABLE bronze.trip\_data\_green\_parquet;

CREATE EXTERNAL TABLE bronze.trip\_data\_green\_parquet

(

Isma sub he column ka datatype set kea ha example: - RatecodeID INT,

extra FLOAT,

)

WITH (

LOCATION = ‘raw/trip\_data\_green\_parquet/\*\*’,

DATA\_SOURCE = nyc\_taxi\_src,

FILE\_FORMAT = parquet\_file\_format

);

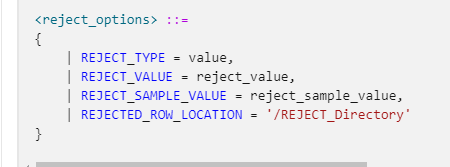
SELECT TOP (100) \*FROM bronze.trip\_data\_green\_parquet;

Handling Rejection

For example, you might want to say that I won’t accept up to 10 invalid records and if there are more then 10 invalid records then query should be failed otherwise, I want to write away the ten invalid record into another file and carry on with the processing

Rejection rows feature is in public preview. please note that rejected rows feature works for delimited (CSV, PARKAE) text files and parser\_version 1.0

reject\_value must be an integer between 0 and 2,147,483,647.



This information about the reject parameters is stored as additional metadata when you create an external table with CREATE EXTERNAL TABLE statement.

IF NOT EXISTS (SELECT \* FROM sys.external\_file\_formats WERE name = ‘csv\_file\_format\_pv1’)

CREATE EXTERNAL FILE FORMAT csv\_file\_format\_pv1

WITH (

FORMAT\_TYPE = DELIMITEDTEXT

FORMAT\_OPTIONS (

FIELD\_TERMINATOR =’,’

, STRING\_DELIMITER = ‘ “ ‘

, FIRST\_ROW = 2

, USE\_TYPE\_DEFAULT = FALSE

, Encoding = ‘UTF8’

, PARSER\_VERSION = ‘1.0’)

);

Bronze.taxi\_zone is a table

IF OBJECT\_ID(‘bronze.taxi\_zone’) IS NOT NULL (what’s that means if table already exists so I’m going to just drop the table)

(In lot of language you get the option to drop the table and it’ll be like drop table if exists and then you can do the table name even SQL server support this but unfortunately, the serverless SQL pole or the dedicated SQL pool don’t support it that’s why we use object\_id.

DROP CREATE EXTERNAL TABLE bronze.taxi\_zone

CREATE EXTERNAL TABLE bronze.taxi\_zone

(

location\_id SMALLINT

brought VARCHAR (15) brought VARCHAR (1) (ERROR CODE)

zone VARCHAR (50)

service\_zone VARCHAR (15)

)

WITH

(

LOCATION = ‘raw/taxi\_zone.csv’,

DATA\_SOURCE = nyc\_taxi\_src,

FILE\_FORMAT = csv\_file\_format\_pv1

REJECT\_VALUE = 10,

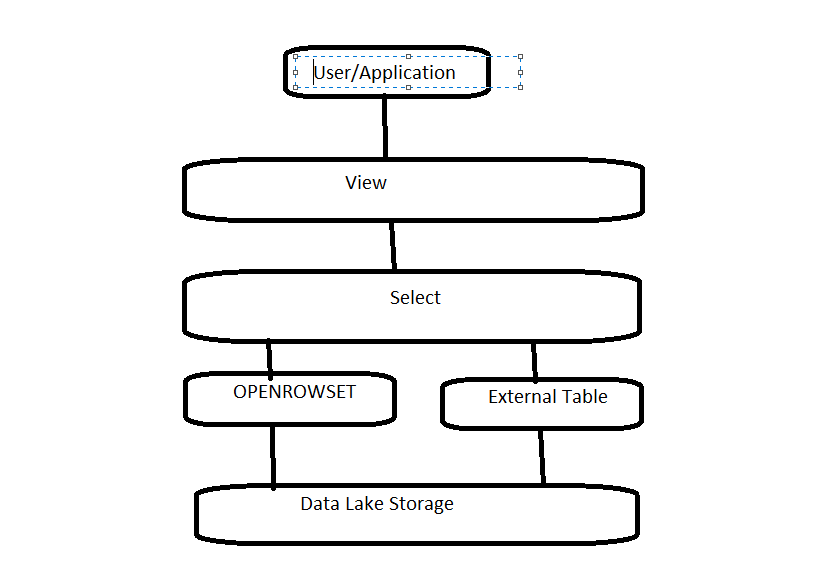
REJECTED\_ROW\_LOCATION = rejections/taxi\_zone

)

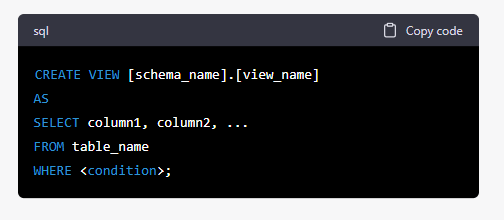
When working with external tables in a database, you may encounter an error message that says "Table already exists." This error occurs when you try to create an external table with the same name as an existing table in the database.

To resolve this issue, you can either choose a different name for your external table or drop the existing table with the same name. Here are the steps to drop the table but when you need to Rember that actually when you dropping a table an external table is not going to drop the data and the data stay at it is it just a metadata over it so we can simply drop the table and recreate it without affecting the data

View



The data is present in the external storage such as ADLs Gen 2, serverless sql can read data from, we know that we can use the openrowset function in the select statement to see the data in the data lake, you can create a view on this select output so that the users and the application can now access the view without worrying about all the parameter that are required for the openrowset function such as storage account ,location, file format, etc also we have to create external table on top of data in data lake storage we can create a view on the select from the external storage too why we should create a view on top of an external table instead of selecting data directly from the external table, an external table on a file has to have all the columns and rows from the data in the data lake storage there is no ability to restrict the data being selected whether it’s row or columns where if you create a view you can select only the columns you want the user to see, as well as apply the where clause to limit data



USE nyc\_taxi\_ldw (for not using the master database)

GO (why we use GO here is that the create view should be the first statement but when we use semi colon so it will be two statement the first one is USE and second one is create view that why we used GO)

DROP VIEW IF EXISTS bronze.vw\_rate\_code (if we again execute CREATE VIEW it gives error that create view already exists that’s why we have to drop view)

CREATE VIEW bronze.vw\_rate\_code

AS

SELECT rate\_code\_id, rate\_code

FROM OPENROWSET (

BULK ‘raw/rate\_code.json’,

DATA\_SOURCE = ‘nyc\_taxi\_src’,

FORMAT = ‘CSV’,

FIELDTERMINATOR – 0x0b

FIELDQUOTE – 0X0b

ROWTERMINATOR – 0X0b

)

WITH

(

jsonDoc NVARCHAR(MAX)

) AS rate\_code

CROSS APPLY OPENROWSET(jsonDoc)

WITH (

rate\_code\_id TINYINT,

rate\_code VARCHAR (20)

)

GO (for another view statement)

SELECT \* FROM bronze.vw\_rate\_code

GO

Partition Pruning

If you have data that’s partitioned in your files, please don’t use external table instead of views with a combination of openrowset and this is the most efficient solution

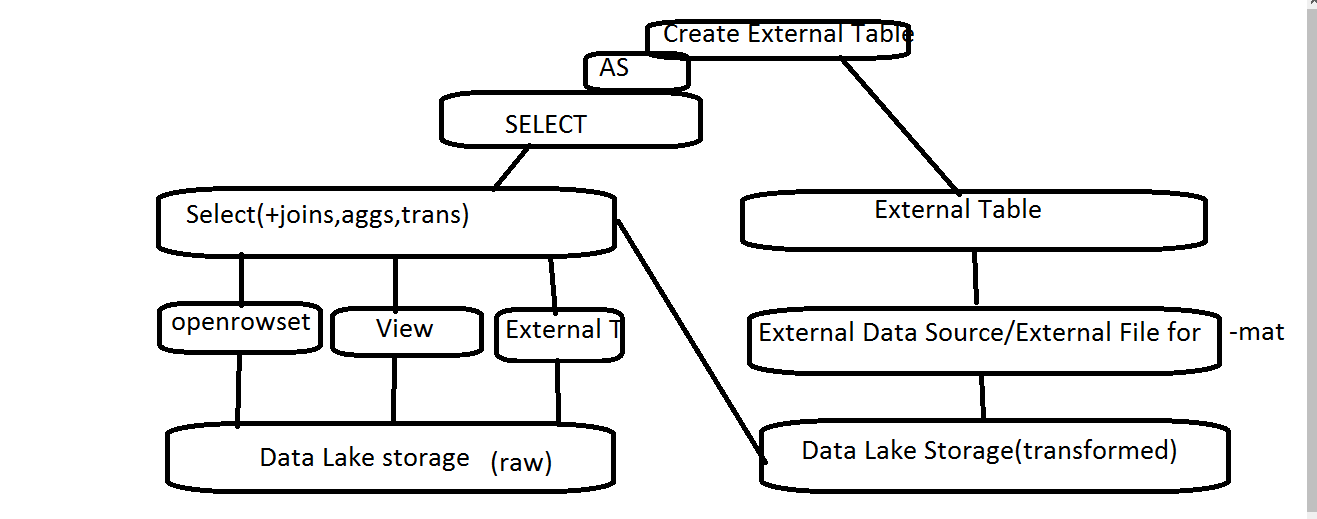
Partition Pruning is an essential performance feature for data warehouse in partition Pruning the optimizer analyzes FROM and WHERE clauses in sql statement to eliminate unneeded partitions hen building the partition access list.

Query multi line json file

Multi line json file are not different from standard json file

Create External Table AS Select (CETAS)

Your aim here to select some data from the storage and transform that data as required and right backed to the storage and also create an external table on the transformed data so that the users can access the transformed data via external tables, we can do all these things through CETAS statement, the create external table part of the Cetas statement will specify the external table definition and the external data source and file format the select part of the Cetas statement will have the logic to select the data from the storage and any function and joins



Here once we have selected the data, we may want to write that data back to the same storage or to the different storage

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas>

Here we create external table as select statement to convert the delimited text file which is CSV format to parquet format, we apply the Cetas and the write the data in the parquet format and create the external table in our silver layer

USE nyc\_taxi\_ldw

GO

IF OBJECT\_ID(‘silver.taxi\_zone’) IS NOT NULL

DROP EXTERNAL TABLE silver.taxi\_zone

GO (we use this because external Tabel already exists) (If you run again then again it will show error which is external Tabel already exists because the Cetas statement is going to write some data in blob storage and the folder already exists in the blob storage and it’s can’t overwrite basically it doesn’t overwrite any data and if there is some data in the blob storage it’s going to fail so we need to delete that data so that the Cetas statement can write the data again so on option Is manually delete the folder)

CRETAE EXTERNAL TABLE silver.taxi\_zone(transformed data present in silver)

WITH

(

DATA\_SOURCE = nyc\_taxi\_src,

LOCATION = ‘silver/taxi\_zone’,

FILE\_FORMAT = parquet\_file\_format ( yaha ya wala aya ga IF NOT EXISTS (SELECT \* FROM sys.external\_file\_format WHERE name = ‘parquet\_file\_format’)

CREATE EXTERNAL FILE FORMAT parquet\_file\_format

WITH (

FORMAT\_TYPE = PARQUET,

DATA\_COMPRESSION = ‘org.apache.io.compress.snappyCodec’

);

)

AS

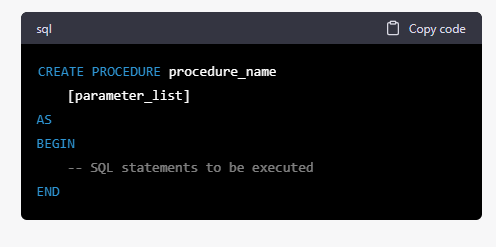
SELECT \*

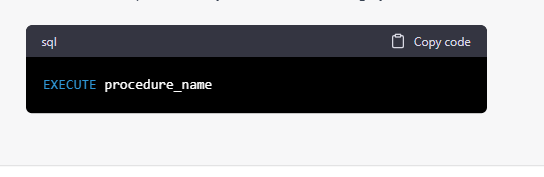
FROM bronze.taxi\_zone;( jo bronze.taxi\_zone jo ke raw format ha who transformed ho kar silver.taxi\_zone ma chala gaya ha)

SELECT \* FROM silver.taxi\_zone;

Stored Procedure

It is basically a group of T-SQL statement stored in a database under a named object, which is called stored procedure. When you execute the stored procedure T-SQL statement will be executed in the order they written in the stored procedure accept input parameters and also, they can return output.



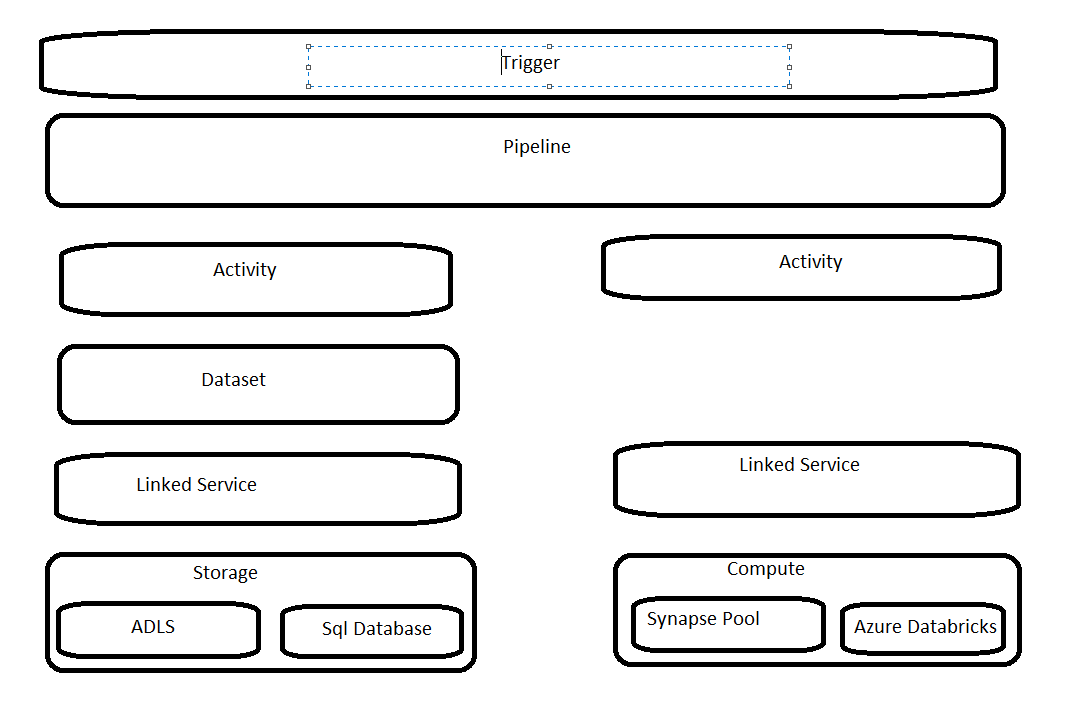
To execute storage procedure

Benefits: -

Resue of code, improve security, Easier Maintenace

Transformed data in parquet file format (partitioned data)

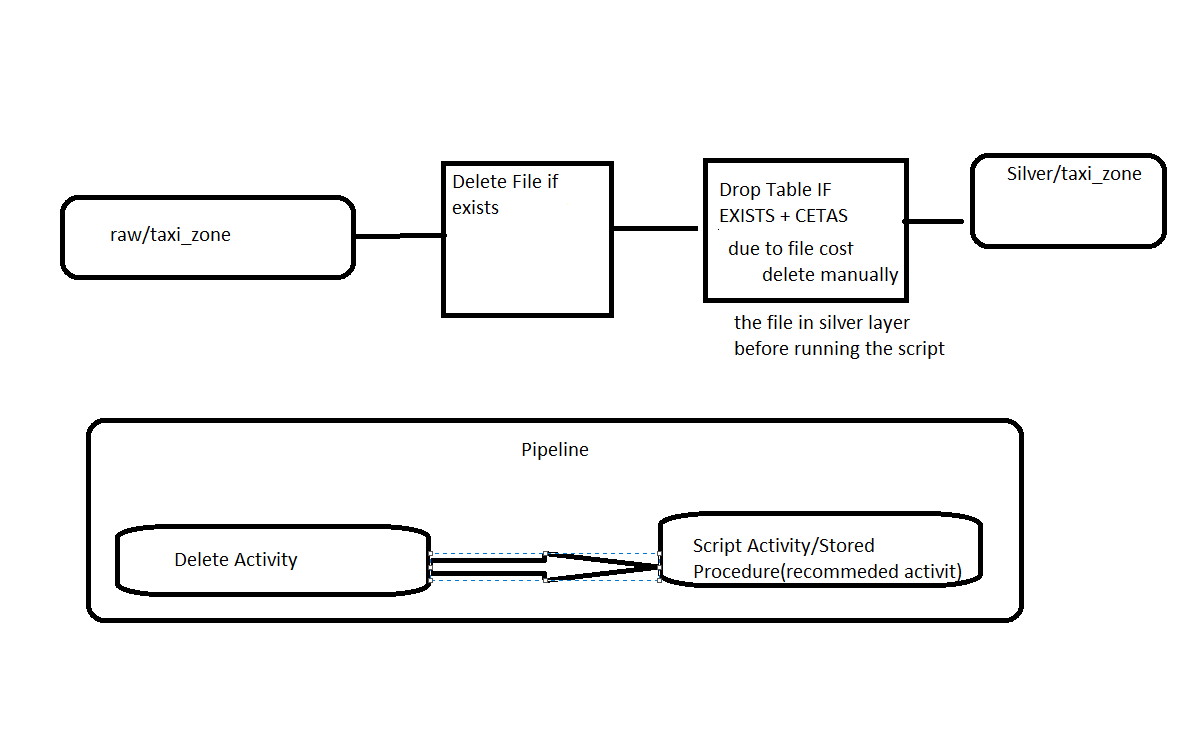
Synapse pipeline component

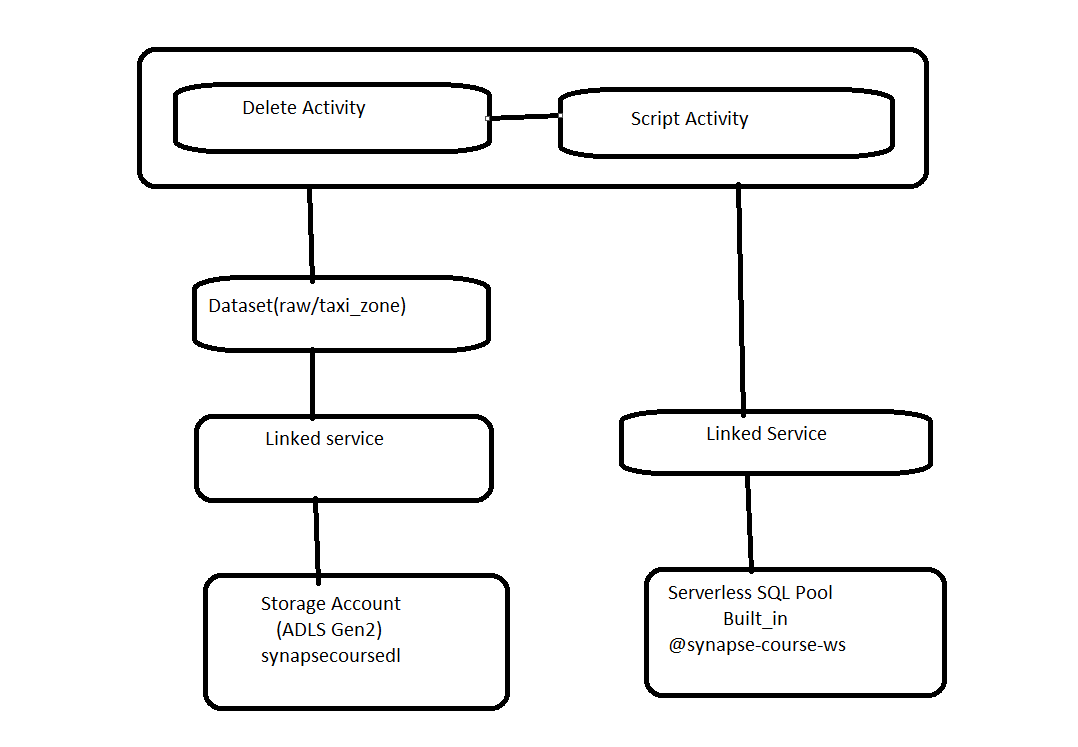


Any kind of action we want to perform in a data project will usually require access to data from some kind of storage such as data lake, cosmos db. etc the other important resource is compute In Azure Synapse Analytics, "compute" refers to the resources and processing power that are used to perform data processing and analysis tasks on large datasets. For example, synapse pool, Azure Databricks. When we create a pipeline in synapse, synapse needs to know how to access this resource let it be storage or the compute for this we need to create a component called linked service it merely have the connection information to one of the resource Link service is a place where you specify the URL of the resource you are trying to access and the credential that provides the access when talking about storage we may want to define the structure of the file the name of the file with in the container etc this kind of information can be stored in another component called dataset, Activity is one simple task which you will define inside the pipeline

Pipeline is logical grouping of activity that together perform a task

What we did to transform the data we simply wrote a script with the CETAS statement to select the data from the CSV file in raw folder and write the data in parquet format in the silver folder but we rerun the script we had to drop the table in order to be able to create it again so to support that we added a statement to drop the table if it already exists even though the issue partially the presence of the file cost another issue so we have to manually delete the file in the silver layer before running the script In our pipeline we would want to automate all of this process which include deleting the file ,running the script with the drop/delete statement and CETAS statement achieve by activity in pipeline.



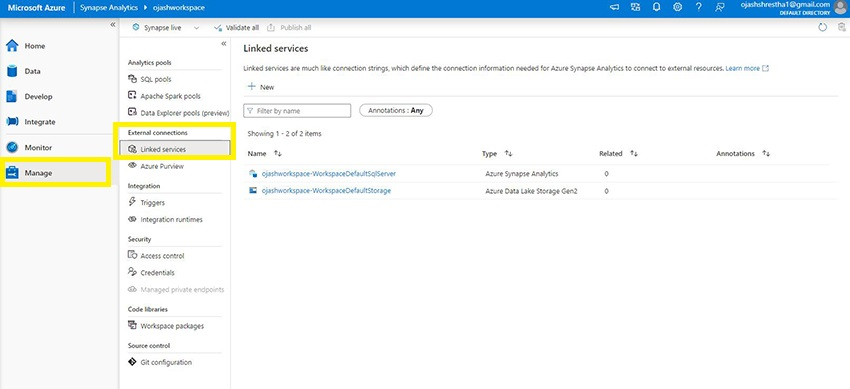


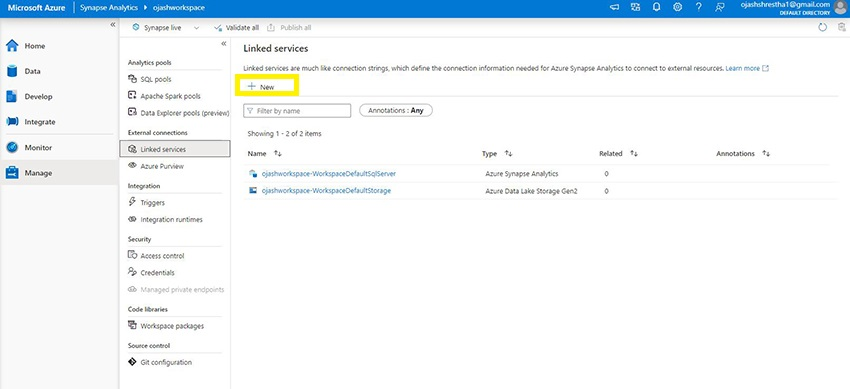
Firstly it need to delete the data from our storage account so we need a linked service which holds the connection information about the storage account and the credentials with the linked services we can connect to the storage account but we still need to inform the delete activity about the folder and the file to be deleted from the storage account for that we need to create a dataset with the required information now when you execute the pipeline the delete activity will be able to connect to the storage account identified the folder to be deleted if exists and delete the folder The script activity need to execute the script in a serverless SQL pool workspace so it needs to know connection information and the credential required to connect to the serverless SQL pool we can create linked service to hold this information now when you execute the script activity it will get the connection information from the link service, connect to the serverless SQL pool an also execute the script.

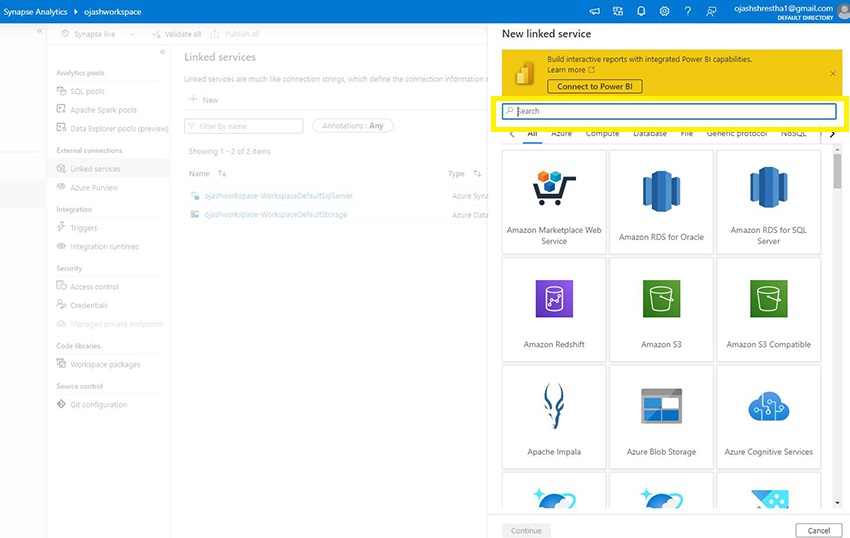
Delete activity = this activity you can run to delete the file if it’s already exists

Script activity = The script activity to run the script, which run the CETAS statement as well as the drop table statement

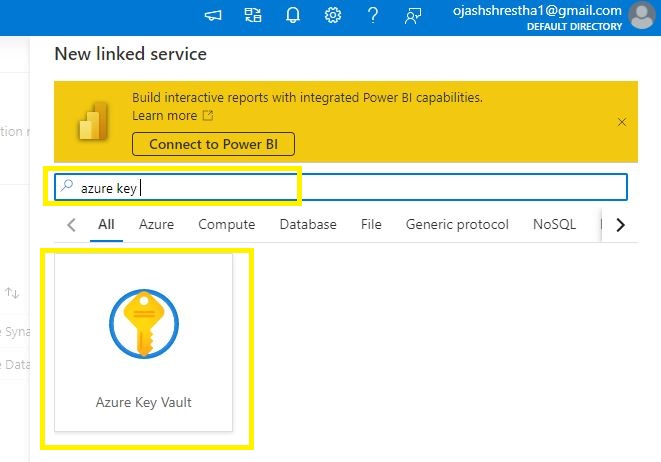
Linked Service

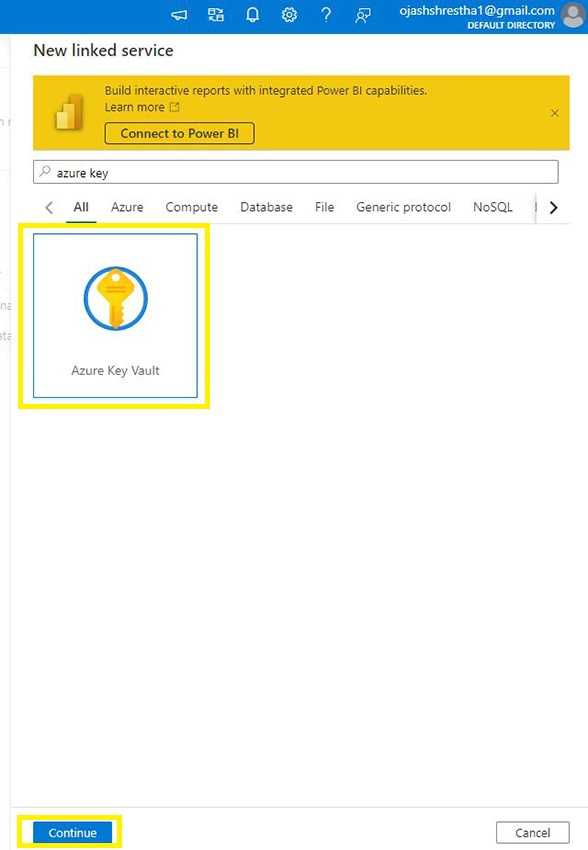


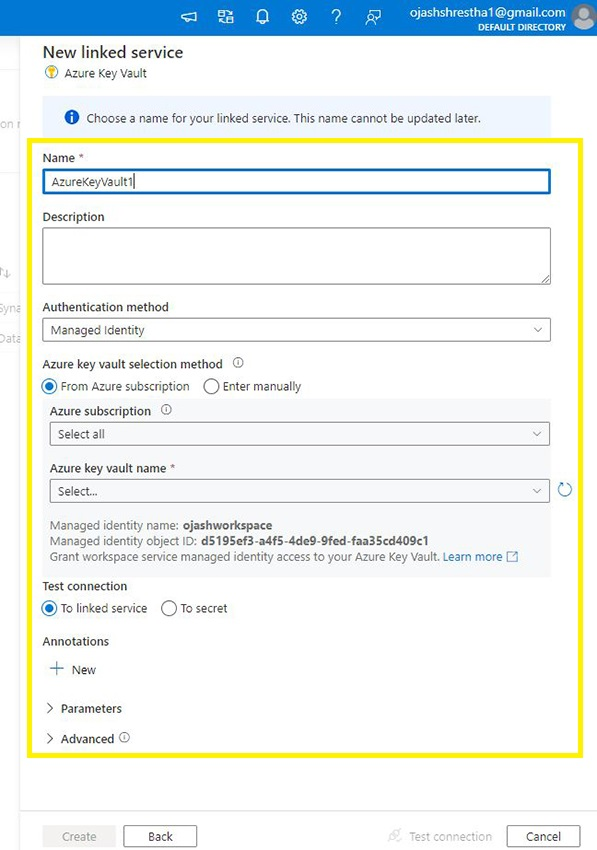


\

Search Azure Data Lake Storage Gen 2



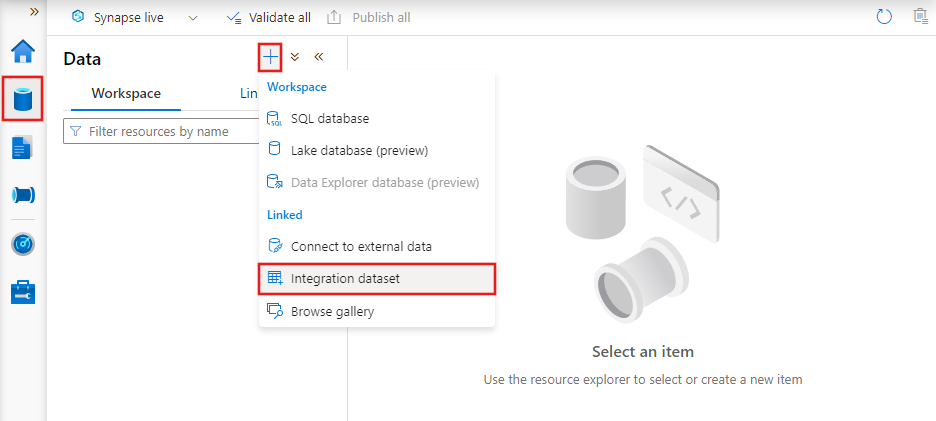




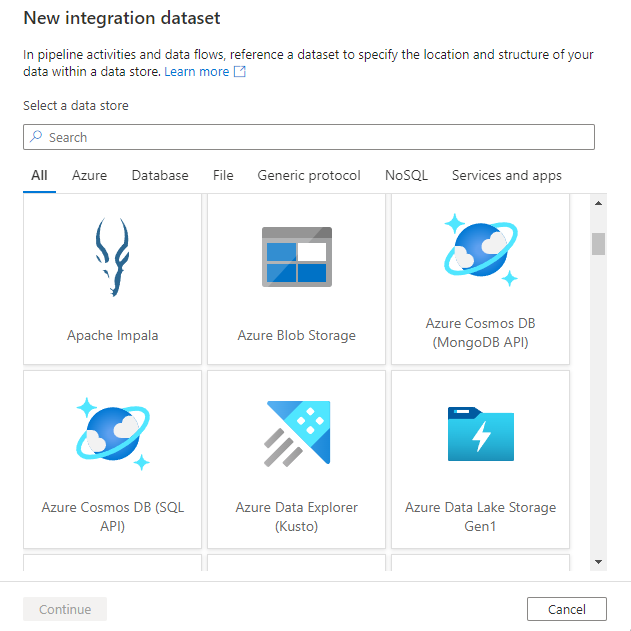
In authentication type choose system assigned managed identity in below you will see managed identity name and that name will be created by default when you create the workspace and that already has access to your storage account as well as the workspace

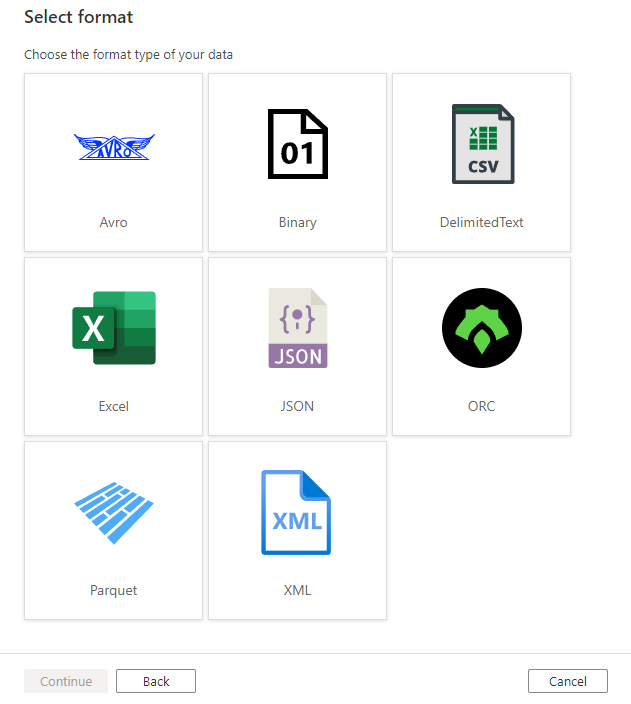
And finally test the connection, linked service only create when you publish them

Create Integration Dataset



Pick Azure Data Lake Storage Gen 2





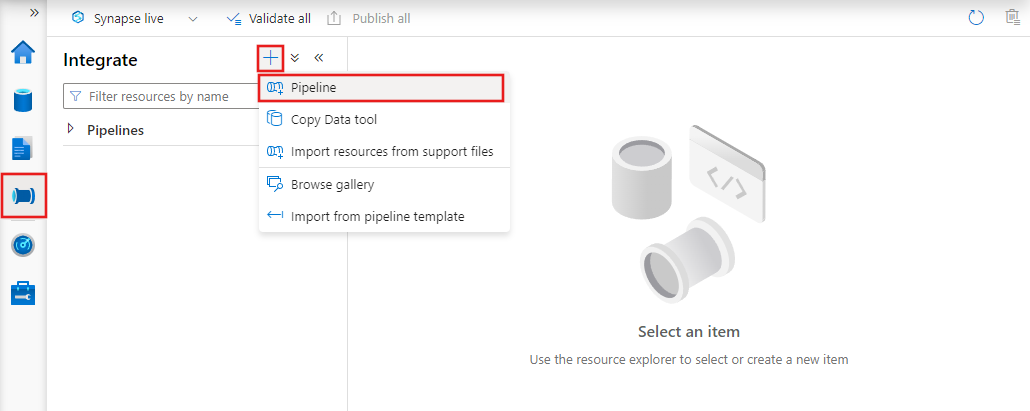
Choose parquet as the format of your data after that it ask Name for your dataset, and we have choose linked service which give access to the dataset or the folder

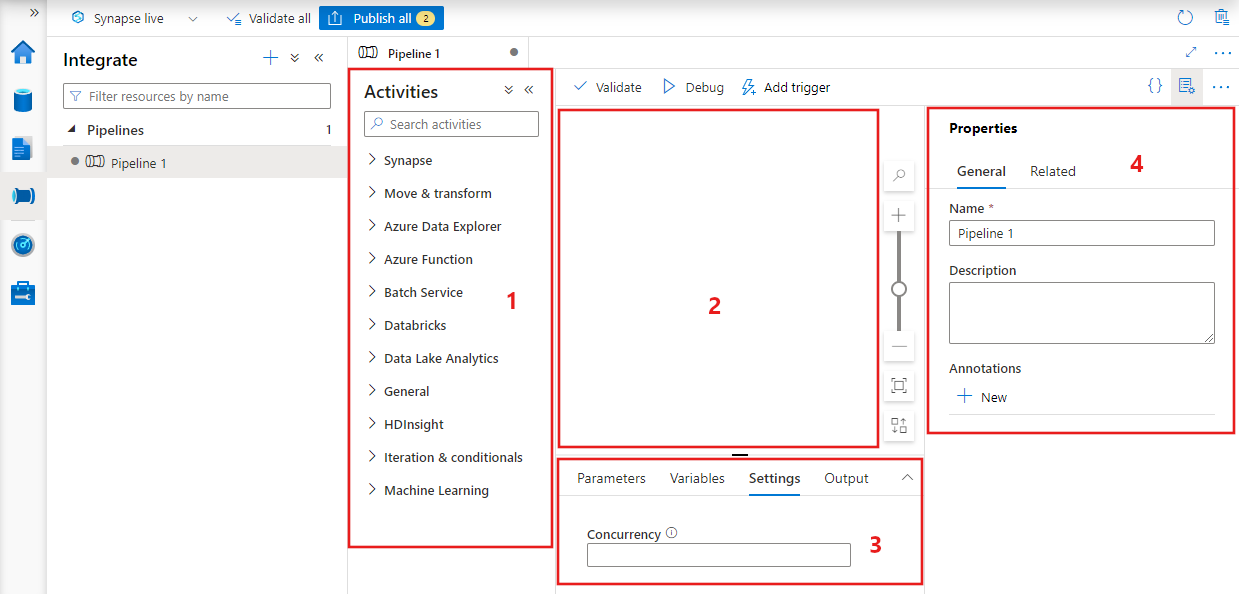


Choose file path which represent dataset, from where you want to select your data and to perform any operation on that data

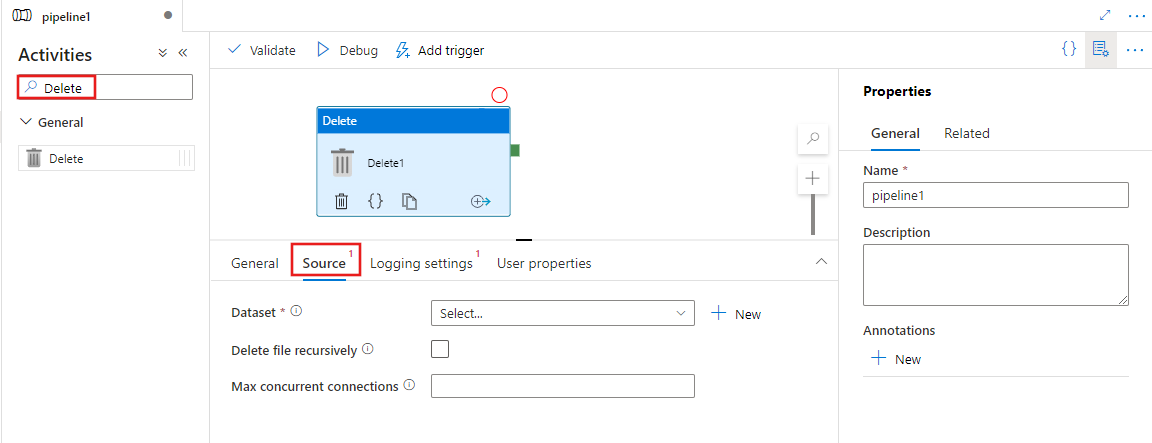
[Datasets - Azure Data Factory & Azure Synapse | Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/concepts-datasets-linked-services?tabs=synapse-analytics)

Create Pipeline

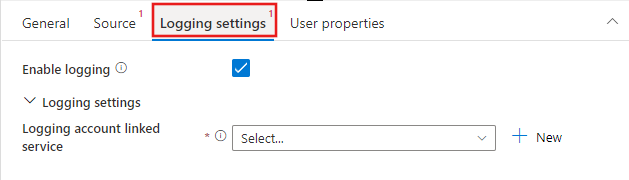




In name section you can give name to your pipeline, Concurrency means that if you want one occurrence of the pipeline to run at a time you set the concurrency to 1(how many pipelines run at a time) as many as you want it to run at the same time you do that



In the above image you can see that it asks the dataset which represent the folder from which we want to delete the data from, so you must specify that dataset



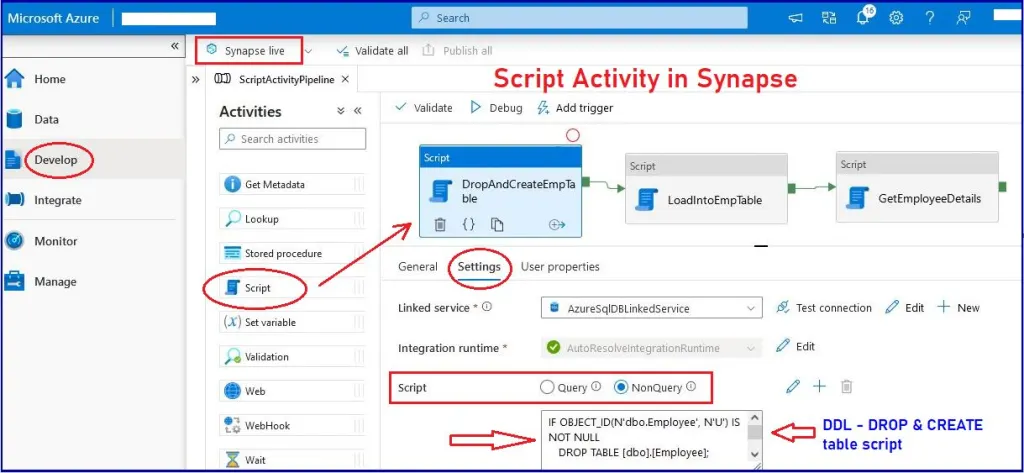
Disable logging

[Delete Activity in Azure Data Factory - Azure Data Factory & Azure Synapse | Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/delete-activity)

Create Script Activity

For that you must create linked service

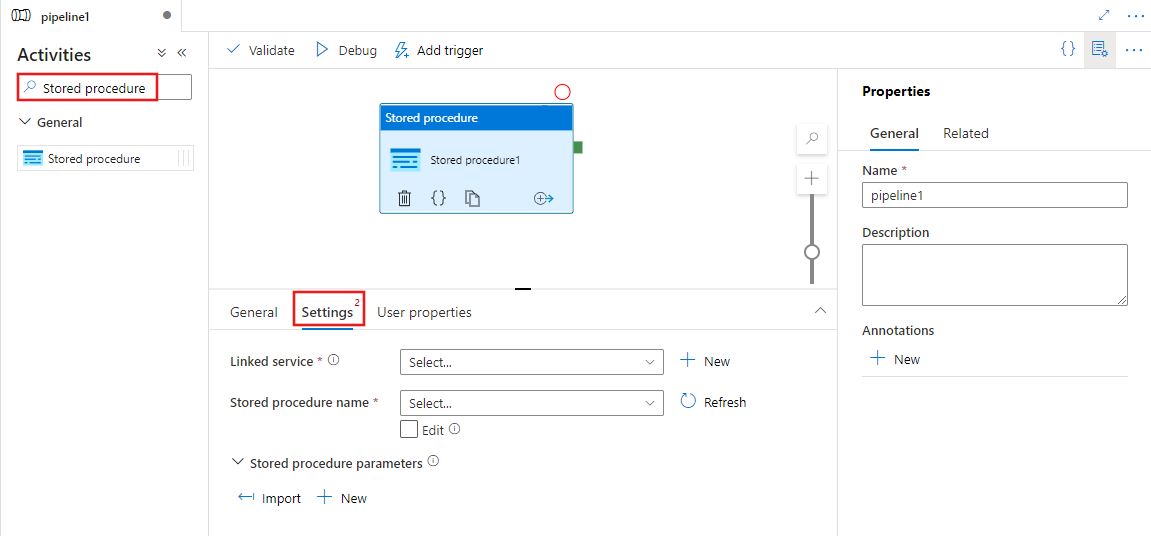
When you create a linked service, it ask you account selection method so if you choose From Azure subscription you can only pick the Dedicated SQL pool so choose Enter manually and when you choose it ask you Fully qualified domain name and you will get that through managed hub. click into the SQL pool in that there is Built-in pool click that you will see the connection string copy from tcp to number which is showed and copy that and placed in the Fully qualified domain name.



If you are SQL is select statement and it return some result and you want to carry that forward, you use the query option otherwise you use the non-query option

Stored procedure activity

Create the linked service



USE nyc\_taxi\_ldw

GO

CREATE OR ALTER PROCEDURE silver.usp\_silver\_taxi\_zone

AS

BEGIN

IF OBJECT\_ID(‘silver.taxi\_zone’) IS NOT NULL

DROP EXTERNAL TABLE silver.taxi\_zone

CRETAE EXTERNAL TABLE silver.taxi\_zone(transformed data present in silver)

WITH

(

DATA\_SOURCE = nyc\_taxi\_src,

LOCATION = ‘silver/taxi\_zone’,

FILE\_FORMAT = parquet\_file\_format

)

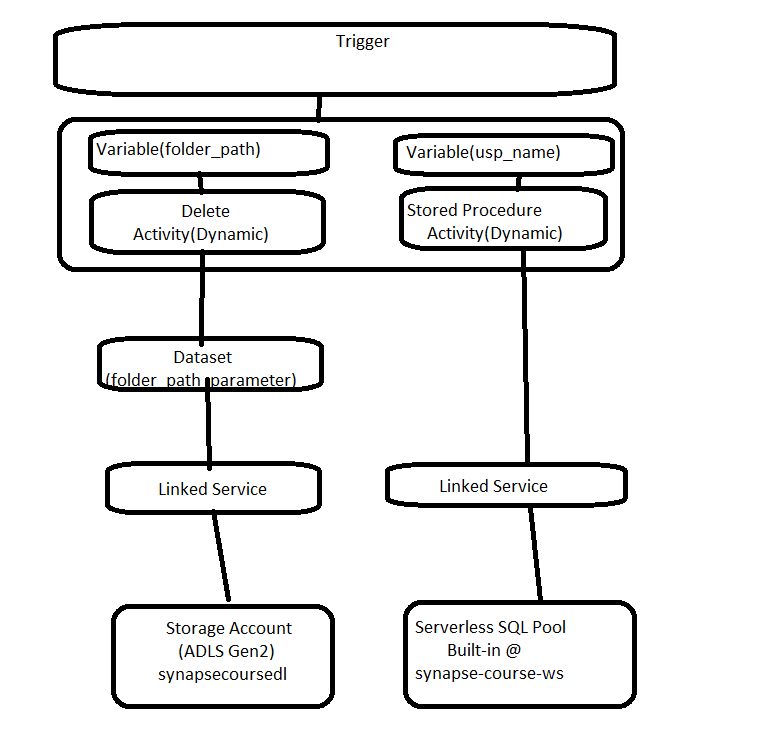
AS

SELECT \*

FROM bronze.taxi\_zone;

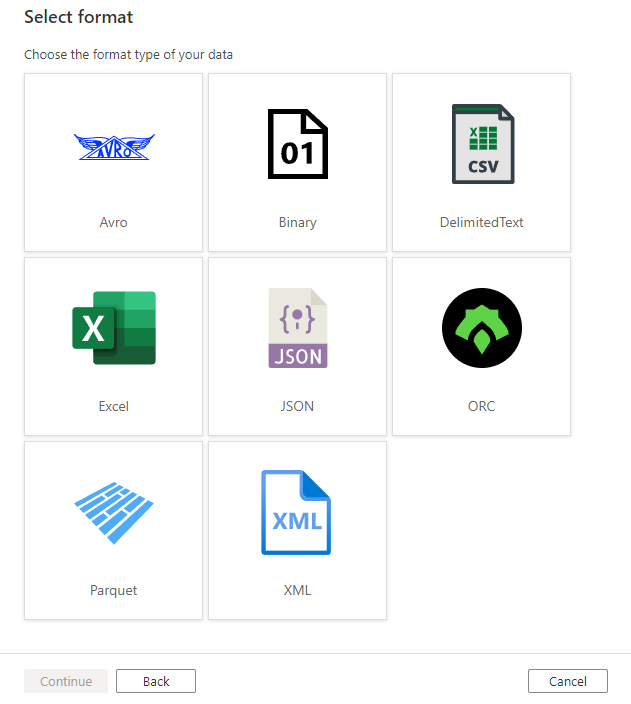
END;

Parameter & Variable



Firstly, we create a dataset with folder path as the parameter, rather than having a dataset pointing to the specific folder. will then add a variable to the pipeline to hold the value of the folder path. So, once we’ve created the variable, we can then pass the value of the folder path variable to the delete activity It can then pass that value to the parameter in the dataset

For taxi\_zone file we have pipeline with two activities (delete, Stored procedure) and dataset similar for calendar also similar we have lot of files, and it is lot to maintain to resolve this situation synapse pipeline provide parameter, parameter available in pipeline, datasets, linked services, Data Flows. via the parameter you can pass values from one component to another. the value of the parameter being sent cannot be changed inside the pipeline or any other component the main benefit of using parameter is that you can reuse the component by sending a different value each time it is being called in that we are also using the variable synapse pipeline offer two types of variable set variable, append variable (isma kay hoga ke hum log variable ma bhajan ga uska Baad delete activity ma jaya ga and then uska bad jaya ga parameter ma jo ke Dataset ma ha)



You create the Dataset, or you can simply copy your old dataset

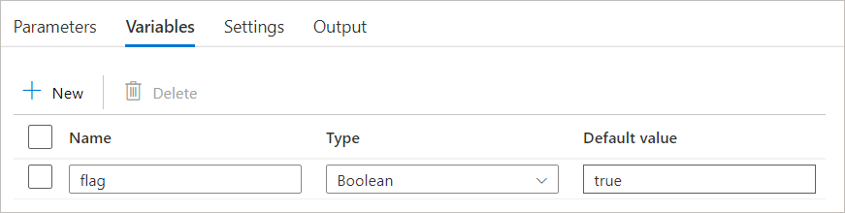
When you choose the parquet format and complete all the thing, after that the dataset is created in that there is parameter give the parameter name p\_folder\_path, type will be string and not give Default Value because we’re going to send value from pipeline to the dataset Now go to connection tab and change file path (Hum log Folder tak jayan ga lakin uska specify nahi karan ga ke Kon sa data) instead of that we specify the parameter which is @dataset().p\_folder\_path ( Here p\_folder\_path) is a parameter. which means the current dataset now go to the Pipeline and click on the **variable** adds the variable which contains the folder\_path and another variable which contain the store\_procedure we have two variable 1) v\_folder\_path 2) v\_usp\_name and give the value of both variable now. we have variable we have use this variable in your activity and for that click on the activity( Delete) and after that click on the source choose the dataset now you see the parameter name which you give in the starting you have to choose the variable for that parameter and for that click on dynamic content and when you click it show all the variable. Now go to the store procedure activity and click on the store procedure name and for that click on add dynamic content to choose second variable and now click on debug.

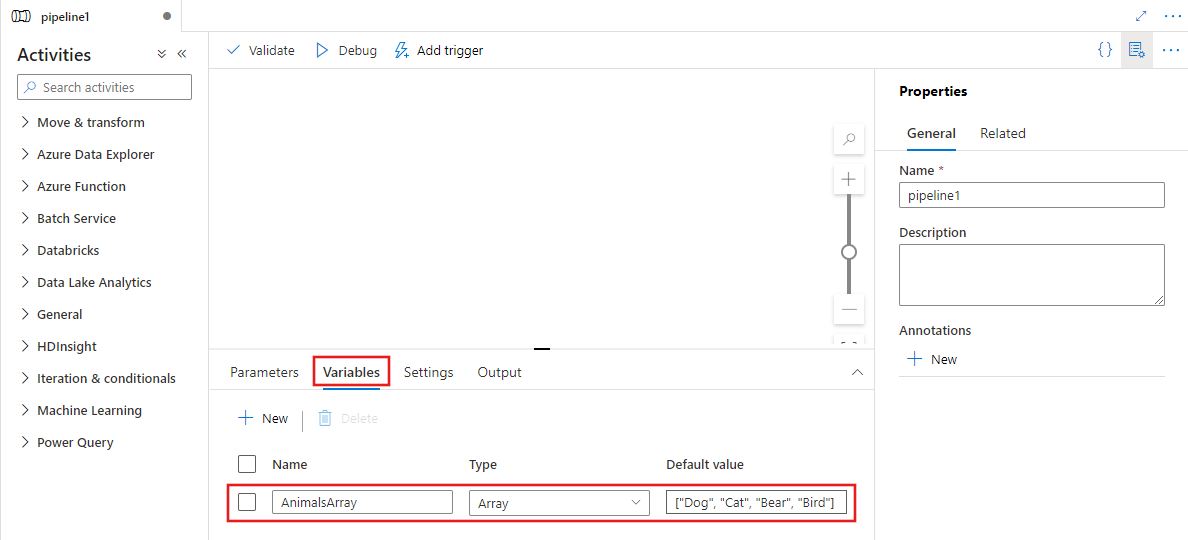
Dynamic Pipeline

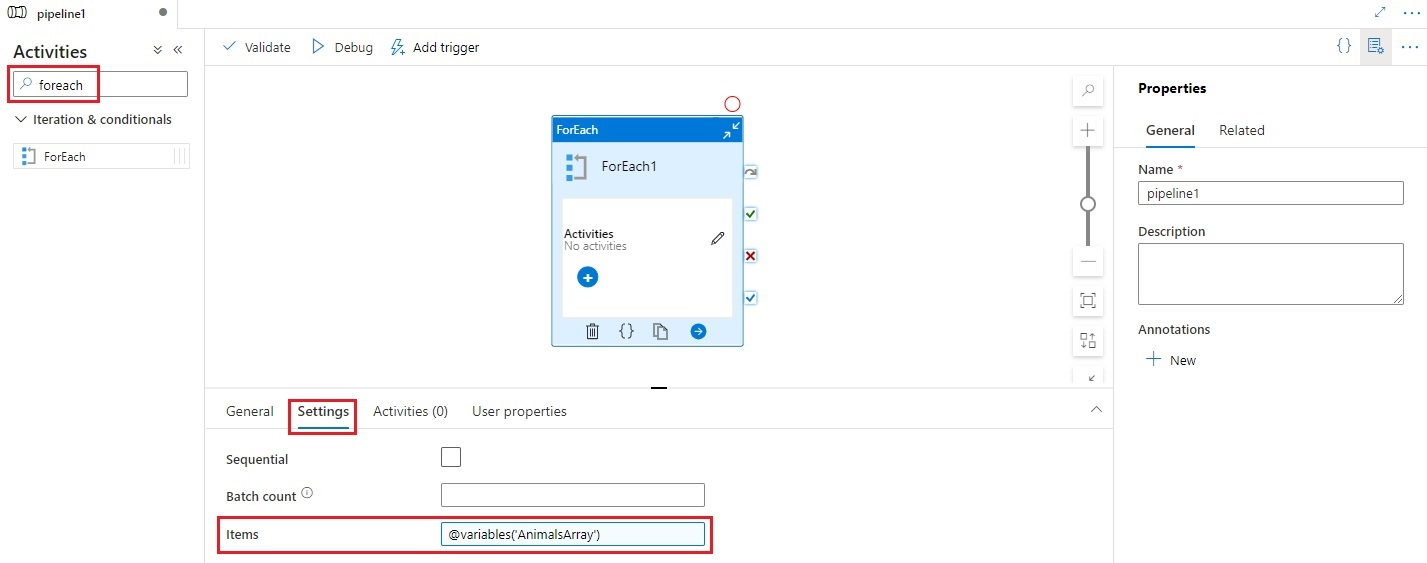
In this solution we use variable of type array, and we use the ForEach Activity to iterate over the array and call the delete and stored procedure activity for each iteration and pass the folder path and stored procedure name for each instance of the array.

[ForEach activity - Azure Data Factory & Azure Synapse | Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/control-flow-for-each-activity)

Go to the variables in that there is type choose array, give name of array and Default value



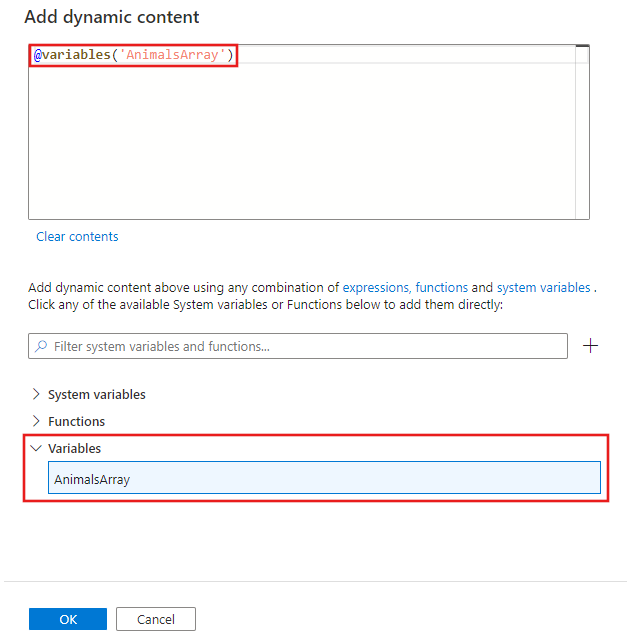
 Here we want to do iteration over the object which we mentioned over the Default value and call both my activity (Delete, store procedure) for each of the iteration.

Select the ForEach activity on the canvas and in **Settings** tab 



There is items section here you want to specify the item so here that variable will come which you mention in the Name section, we are going to iterate over the object

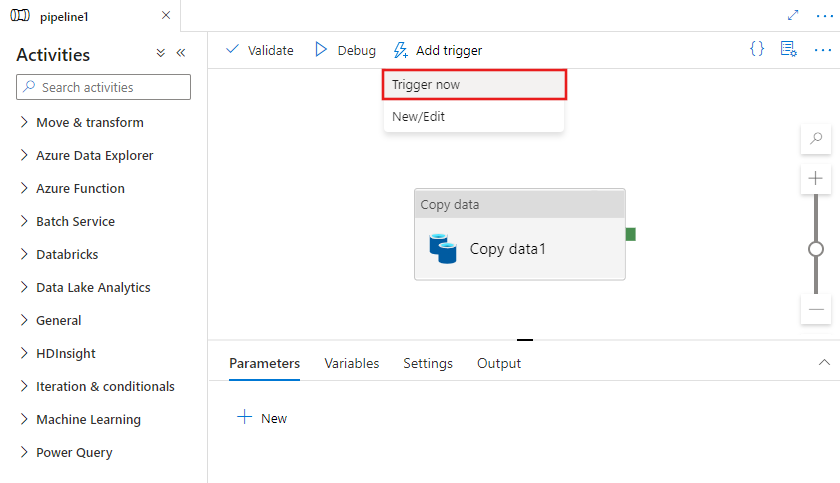
Click on Add dynamic content and choose the variable name you are given in the starting before choosing the foreach activity



When you iterate over the variable you want to perform certain activity so there are two activity delete and stored procedure so go to the activities section of foreach activity and select both the activity and copy them and past in the foreach activity. In both the activity we must refer to the array object rather than the variable itself in order to do that Go to the Delete activity which is present in foreach activity and go to the source section of Delete activity and now for the parameter we choose the ForEach iterator and for stored procedure name we choose the ForEach iterator.

App ko har ek ka lea stored procedure create karna Hoga

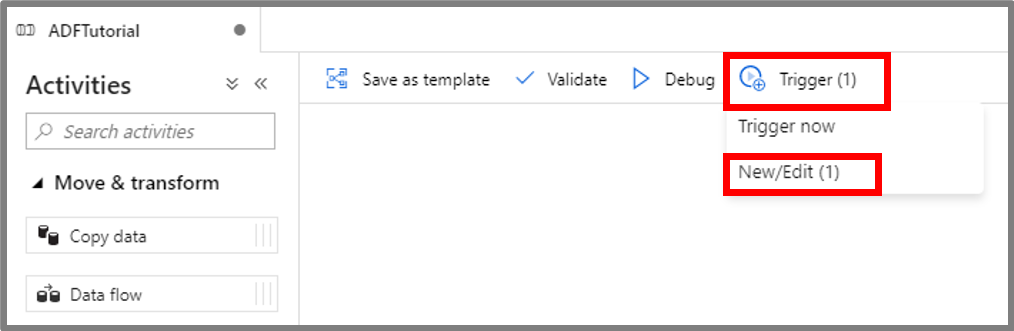
Manual trigger

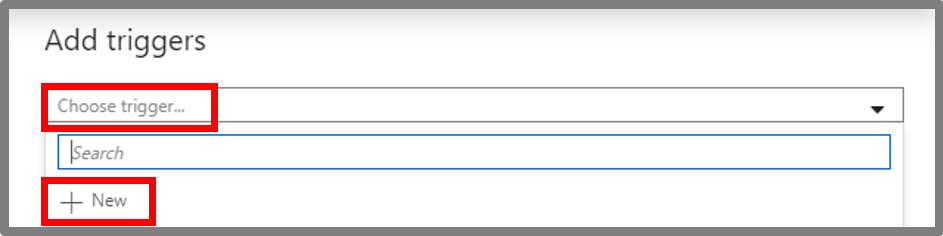


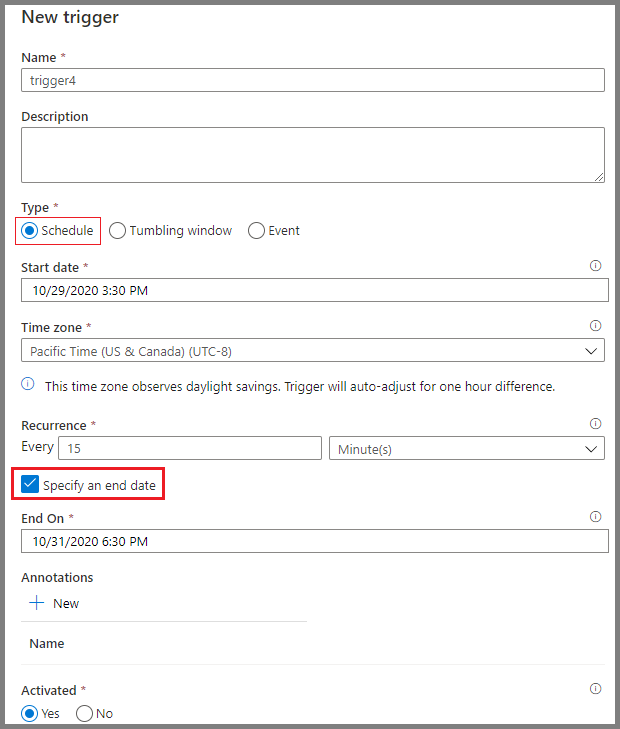
When you go on Add trigger and click on Trigger now so that will trigger an on demand run so it’s not going to be an automated run, it’s going to be an on-demand run click on OK. To see the trigger, go to the monitor in that there is an integration section in that there is two run one is pipeline run and next one is trigger run so trigger run you see when the trigger was executed, and the pipeline run will be when the pipeline came into running but we just did a manual run so you will find the pipeline run(samaj lo ke manual trigger under pipeline run ma rahta ha) under the pipeline run

Automated Trigger

[Create schedule triggers - Azure Data Factory & Azure Synapse | Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/how-to-create-schedule-trigger?tabs=data-factory)







Four Type of trigger Schedule, Tumbling window, Storage events, Custom events

Schedule = Scheduled triggers basically run on the wall clock you specify when you want your trigger to start and then how often do you want trigger to run

Tumbling window = Tumbling window trigger a run on an interval, so you specify your start and end and then it’ll repeat the start and end ex: - if you wanted to grab for every 6 hours that you specify, you are interval 6 hours and then you specify the start date there are no overlaps allowed, so it’ll have to be exact 6 hours period to be repeating that

Storage event = Storage event trigger works on the azure storage, so that can be based on blob, or a container being created or deleted so when you create a blob or you delete a blob, then you can invoke a pipeline from that.

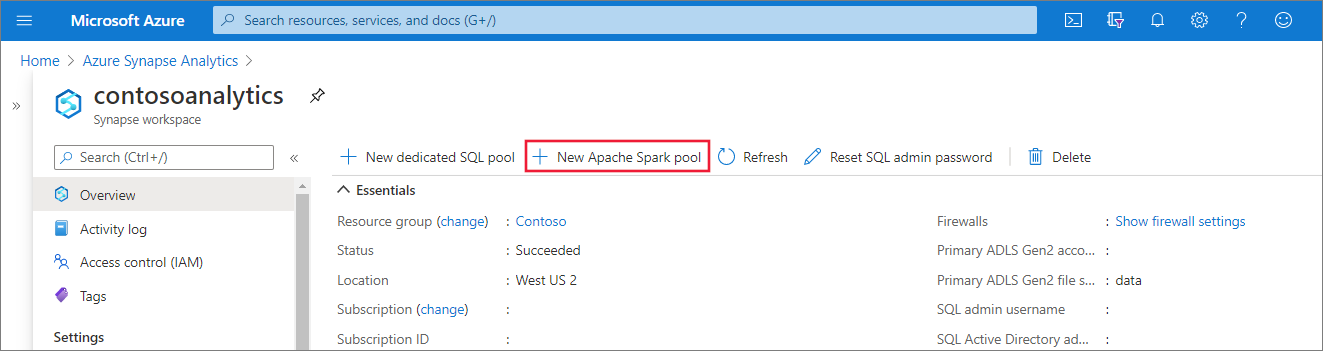
Custom event = Custom events works based on events being pushed into event grid so the synapse pipeline can subscribe to the topics in your event grids and when an event arrives, you can trigger the pipeline based on that

Spark Pool

In the Synapse workspace where you want to create the Apache Spark pool, select **New Apache Spark pool**.

[Quickstart: Create a serverless Apache Spark pool using the Azure portal - Azure Synapse Analytics | Microsoft Learn](https://learn.microsoft.com/en-us/azure/synapse-analytics/quickstart-create-apache-spark-pool-portal)

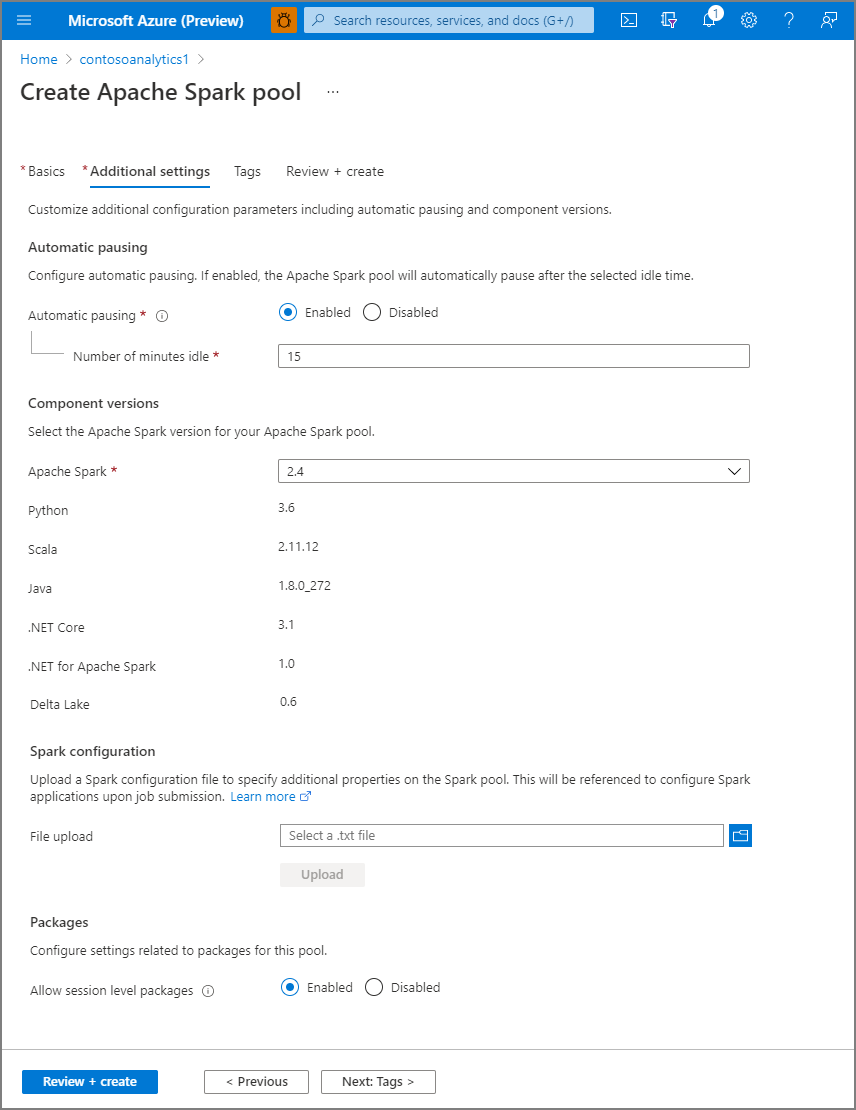
[How to use Synapse notebooks - Azure Synapse Analytics | Microsoft Learn](https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-development-using-notebooks)



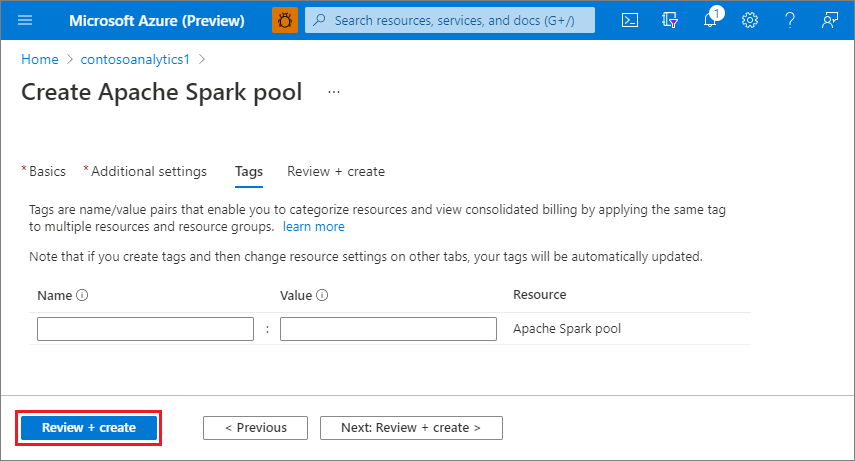


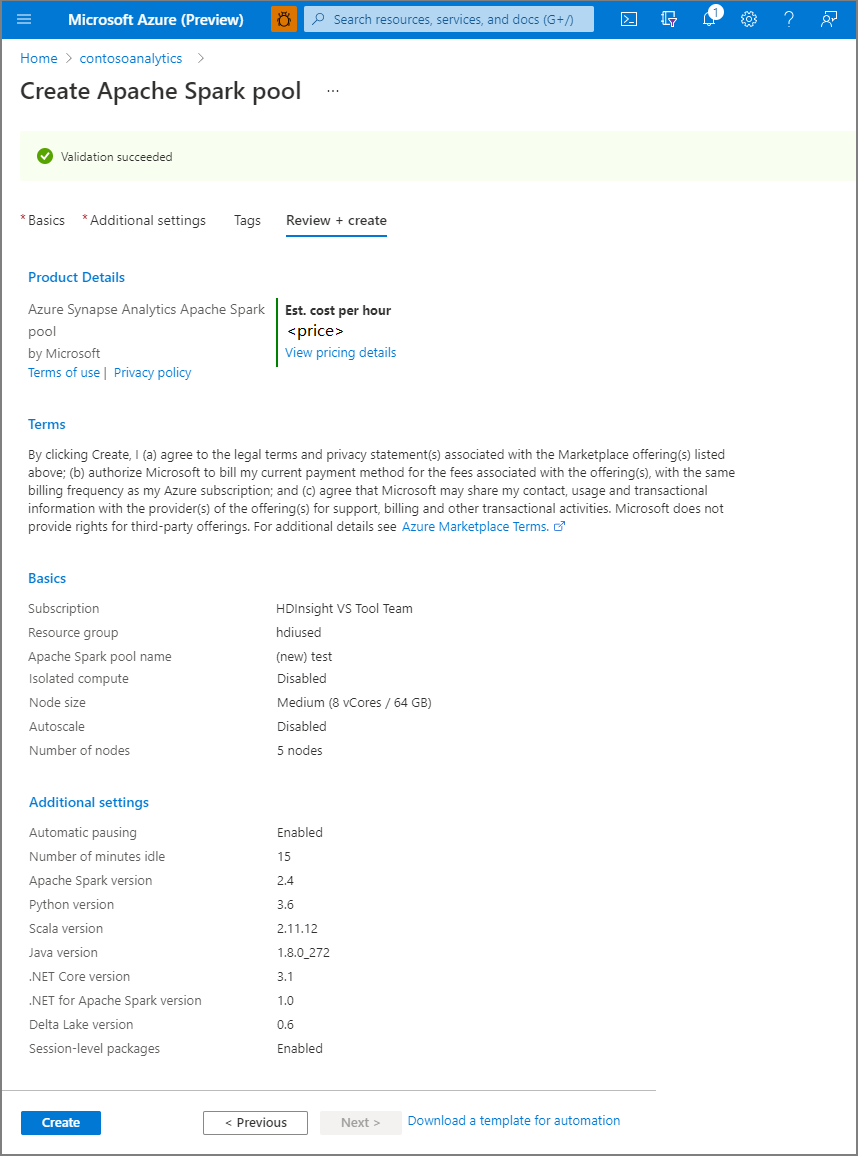
Give name of Apache spark pool, Node size, you can specify minim number of node and max number of node and synapse scale your pool (number of node) depending on your workload

|  |  |  |
| --- | --- | --- |
| **Size** | **vCore** | **Memory** |
| Small | 4 | 32 GB |
| Medium | 8 | 64 GB |
| Large | 16 | 128 GB |
| XLarge | 32 | 256 GB |
| XXLarge | 64 | 512 GB |
| XXX Large (Isolated Compute) | 80 | 504 GB |



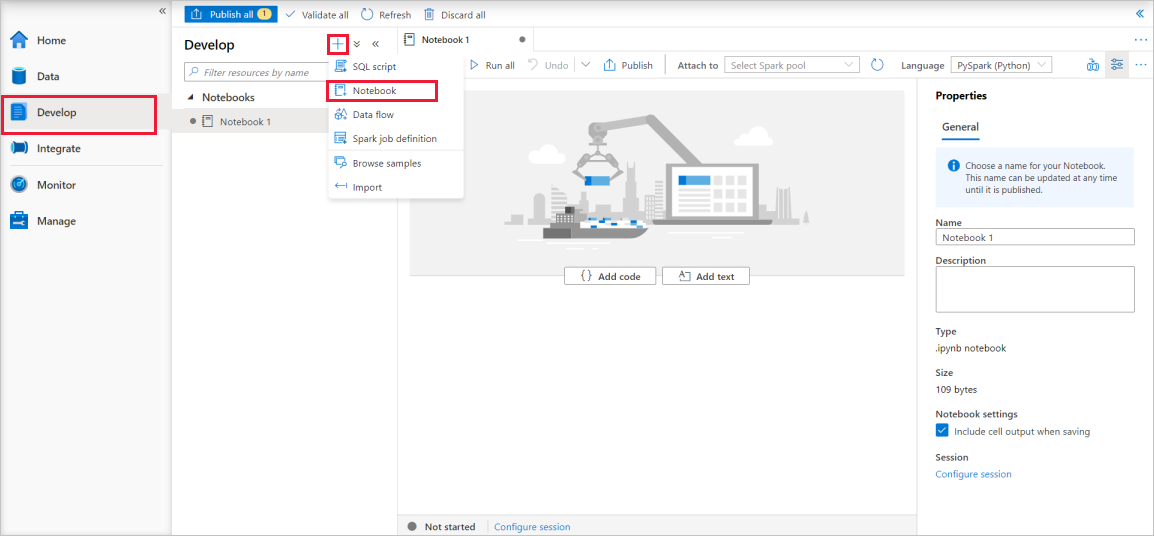
Number of minute idle which is 15 and its tell that the spark pool will be shut- down after 15 min of inactivity





Notebook

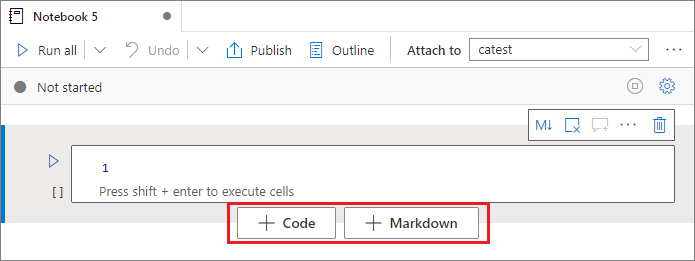
In order to interact with the spark pool, you will have to create a notebook

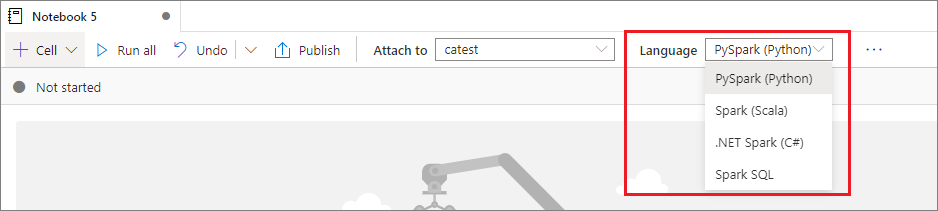


Attach to means attach your notebook with spark pool

Synapse notebooks support four Apache Spark languages:

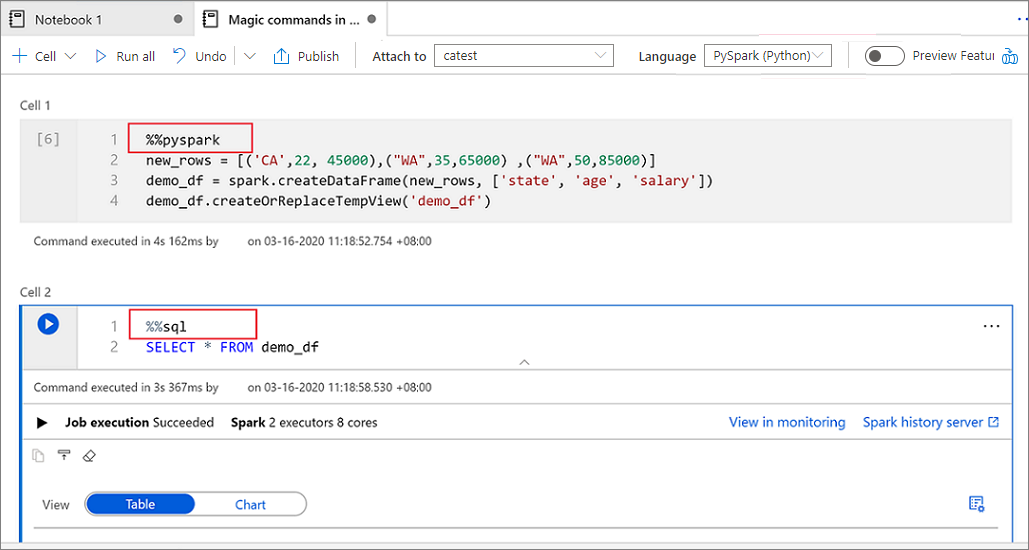
* PySpark (Python)
* Spark (Scala)
* Spark SQL
* .NET Spark (C#)
* SparkR (R)





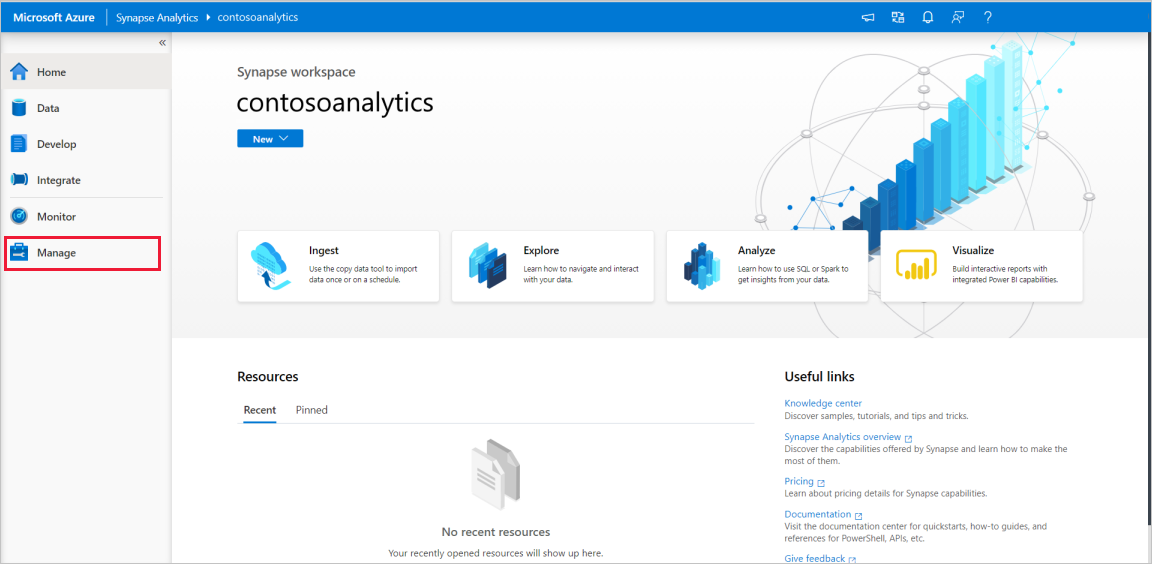
You can use multiple language in one notebook

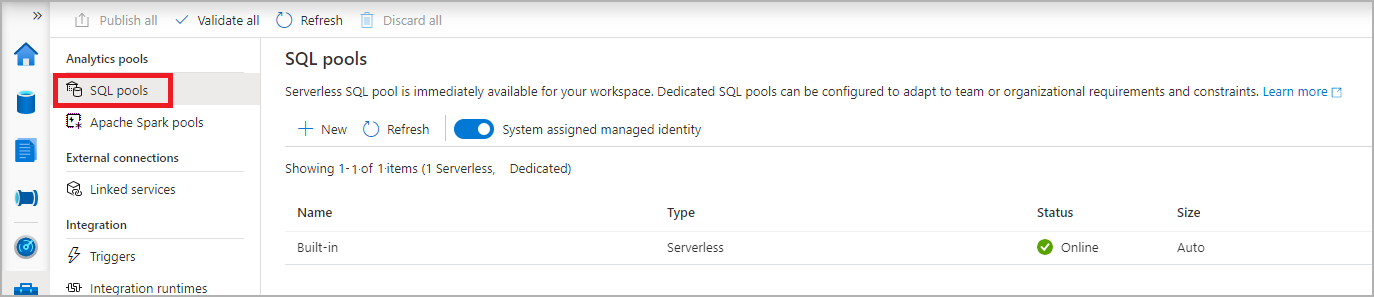
|  |  |  |
| --- | --- | --- |
| %%pyspark | Python | Execute a **Python** query against Spark Context. |
| %%spark | Scala | Execute a **Scala** query against Spark Context. |
| %%sql | SparkSQL | Execute a **SparkSQL** query against Spark Context. |
| %%csharp | .NET for Spark C# | Execute a **.NET for Spark C#** query against Spark Context. |
| %%sparkr | R | Execute a **R** query against Spark Context. |

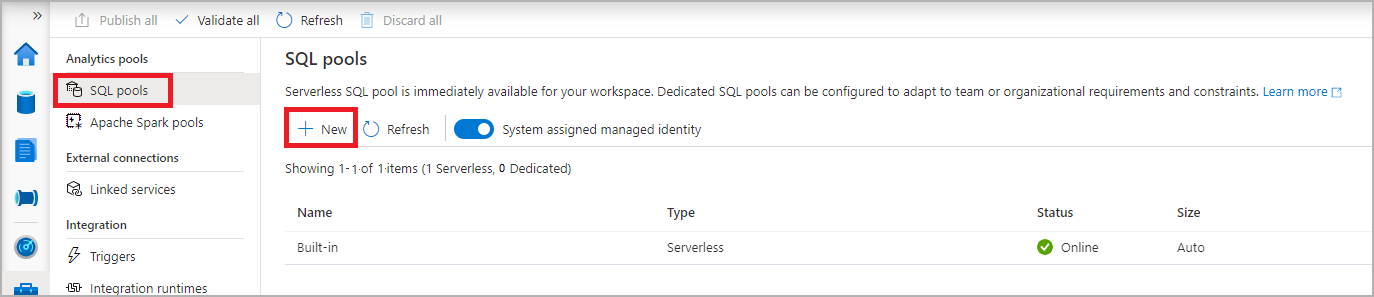


Dedicated SQL Pool

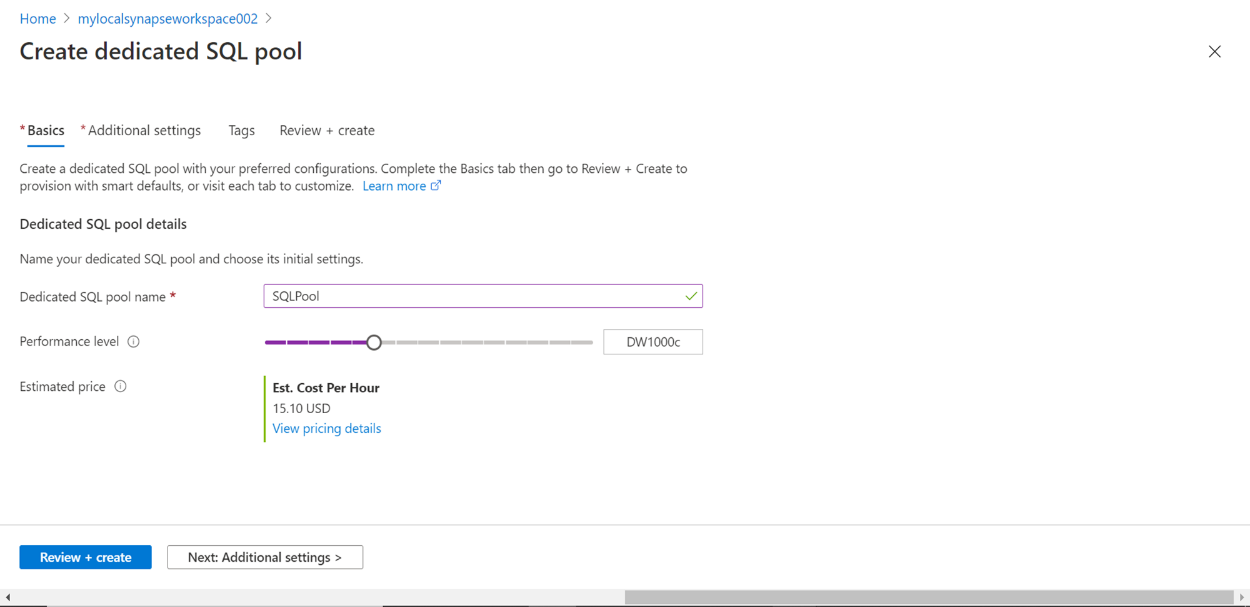
Dedicated sequel pool is basically a distributed query engine that you can use to perform high performance big data analytics using the familiar T-SQL query language unlike serverless SQL pool, dedicated SQL pool also offers a data storage solution that stores data in a table structure with columnar format so it’s basically a traditional data warehouse with a massively parallel processing engine some of you might already be familiar with its old name as azure SQL Dataware house incorporated into the synapse as dedicated SQL pool

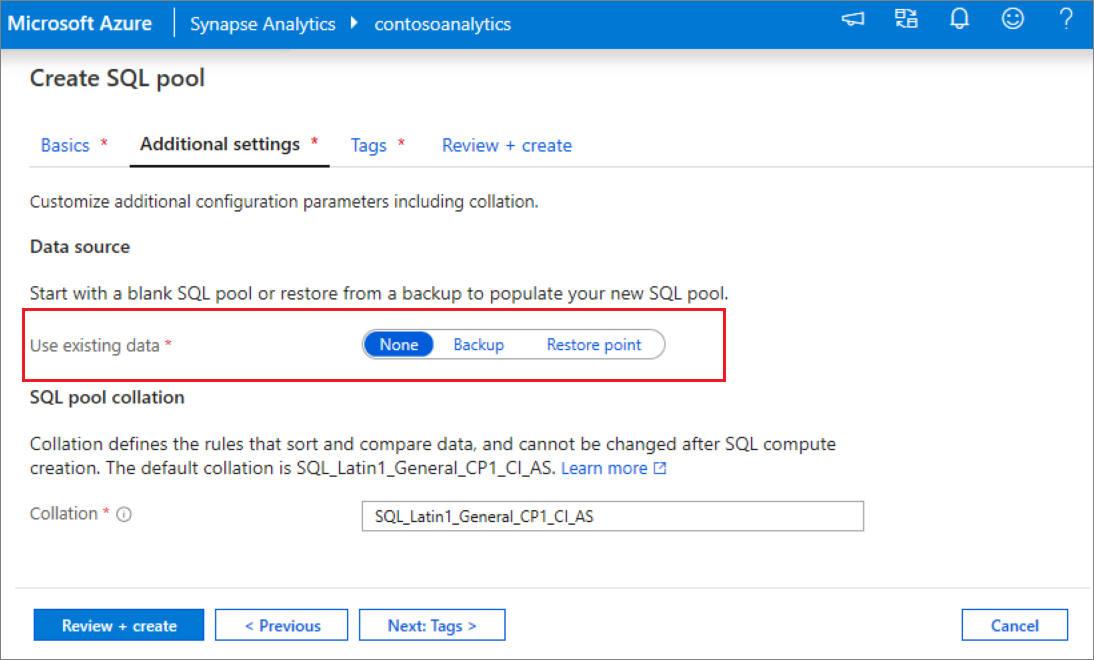


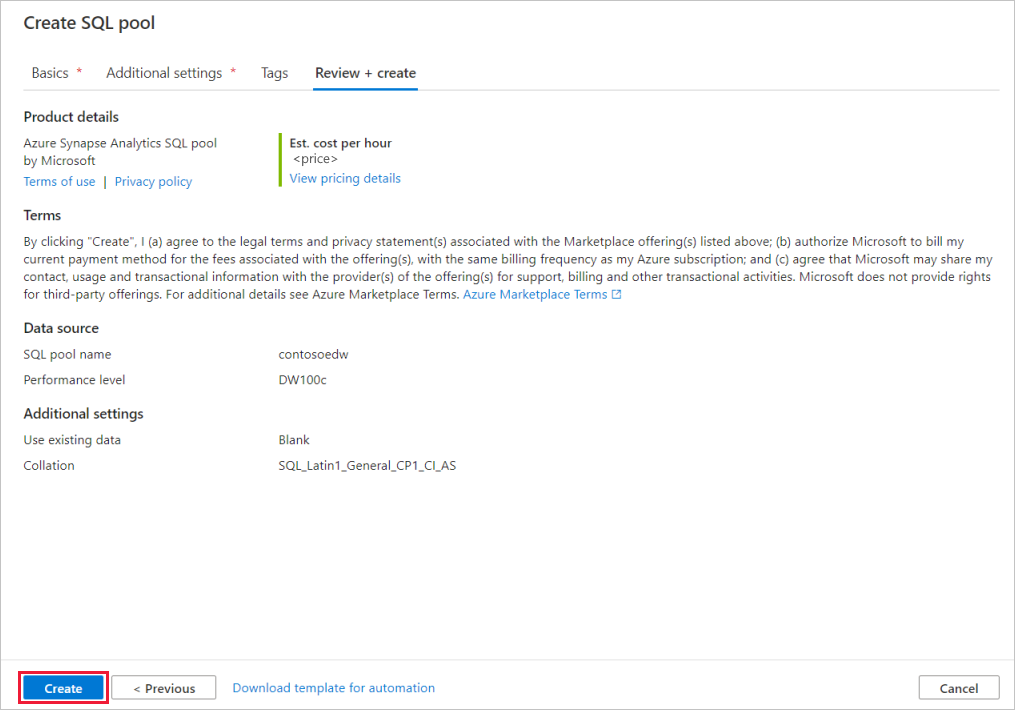




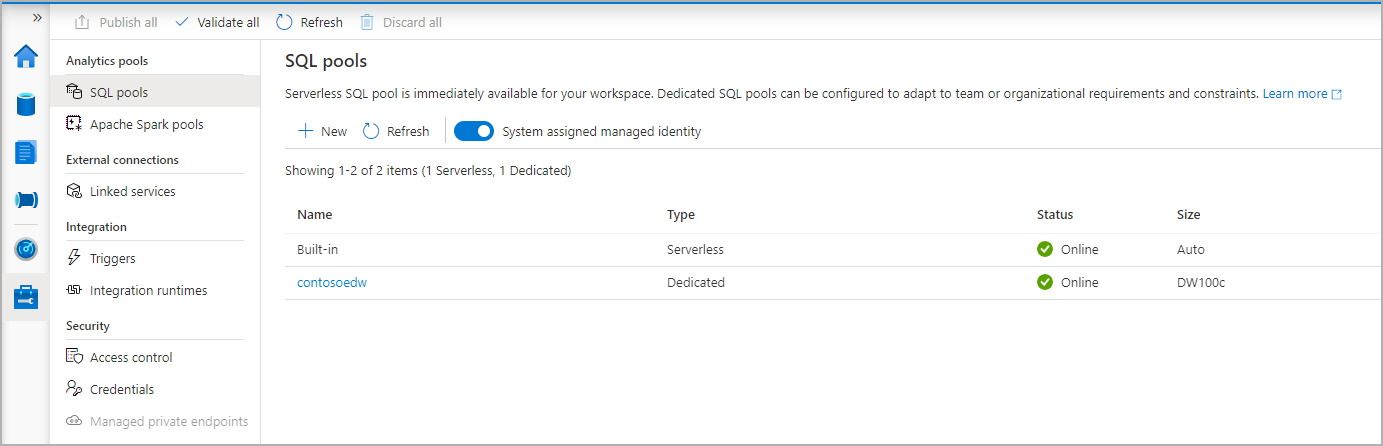
|  |  |  |
| --- | --- | --- |
| **Setting** | **Suggested value** | **Description** |
| **Dedicated SQL pool name** | Any valid name | Name of the dedicated SQL pool. |
| **Performance level** | DW100c | Set to the smallest size to reduce costs for this quickstart |





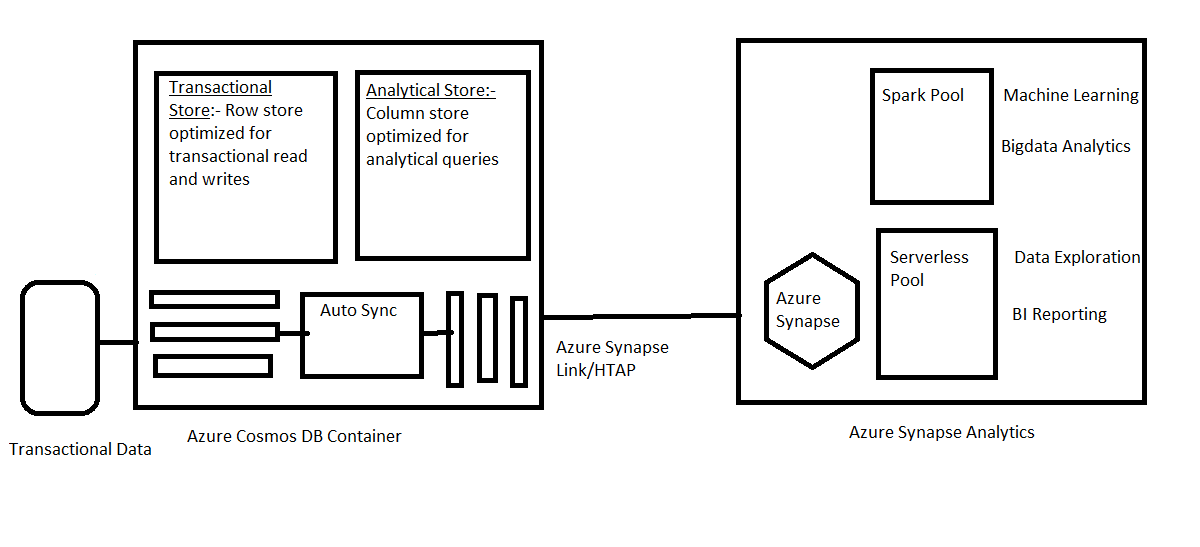


After the provisioning is completed, navigating back to the workspace will show a new entry for the newly created SQL pool.



As you see that the dedicated SQL pool has been provisioned and it’s online and it’s available for us to use when it’s online it means two thing one it’s been deployed and it’s available for us to use and second you are being charged at this point in time for the service so that’s the main difference between the serverless SQL portal, which as you see that serverless SQL pool is online you are not being charged at all ,actually you charged only when you processing any data but the dedicated SQL pool is a dedicated resource which means is you are going to be charged while that’s online.

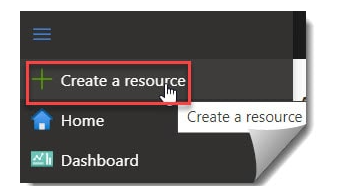
Synapse Link for Cosmos DB

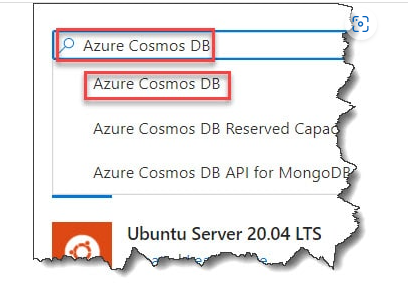
 Let’s assume that your transactional system is using azure cosmos DB to serve to the business application to support the needs of the transactional processing application the data is stored in a transactional store the data is stored as a resource in the cosmos DB container so that the data stored is optimized for transactional read and write, increasingly business are interested in this data to perform big data analytics and machine learning in near-real-time ,the traditional approach of using ETL introduces latency to the data as well as complexity to the data pipeline this where synapse link came into play when you enable synapse link the data in the transaction store will be replicated to the analytical store and data in the analytical will be in columnar format which is optimizes for querying via analytical query engine. The replication from transactional store to the analytical store is carried out by an auto sync process, which is gully managed by azure, we now need to access this data from synapse in order to that we need to create a link service in synapse to the cosmos DB database via the link service we can either use the spark pool or the serverless SQL pool to interact with cosmos DB as we know we use spark pool to do ML or bid data and use serverless to perform data exploration and BI reporting.

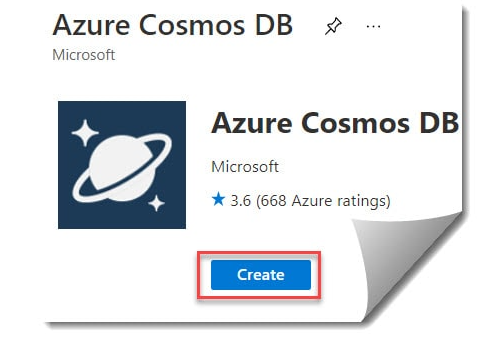
Synapse Link

In Azure Data, a synapse link is a connection between two Azure resources: Azure Synapse Analytics and Azure Data Services. The synapse link allows you to move data seamlessly between these two services. When you create a synapse link, you can access and transfer data from your Azure Synapse Analytics workspace to various Azure data services like Azure Blob Storage, Azure Data Lake Storage, Azure Cosmos DB, Azure Event Hubs, and Azure SQL Database.

Create Cosmos DB

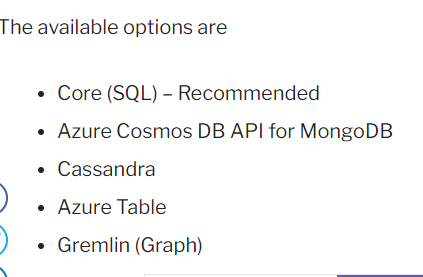


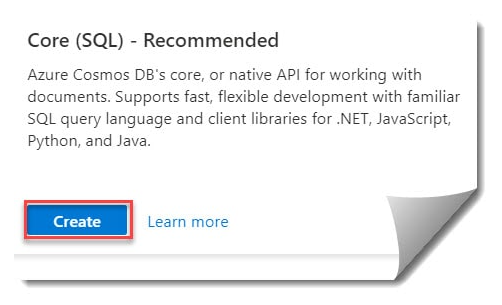


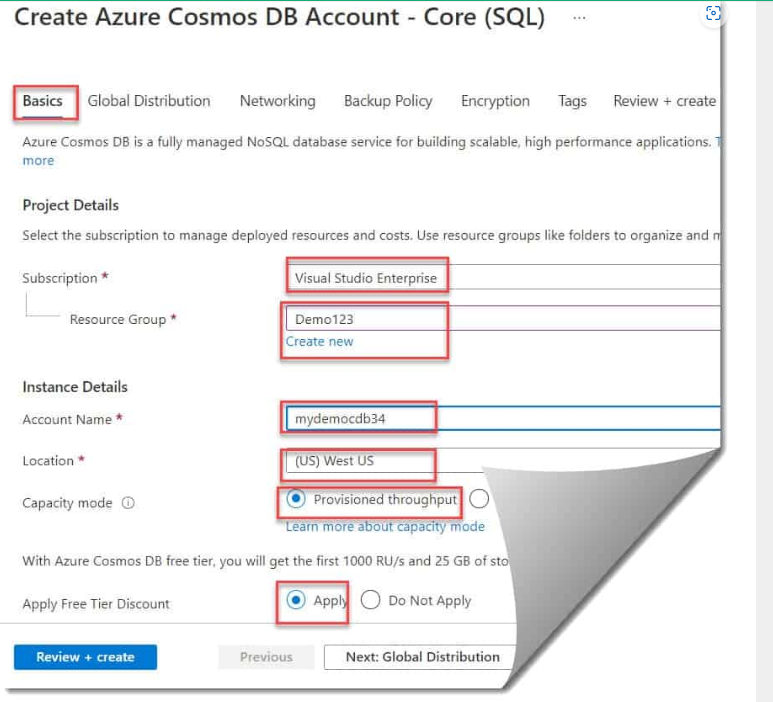


On the Create Azure Cosmos DB Account – Core (SQL) page, provide the below details on the Basics tab.

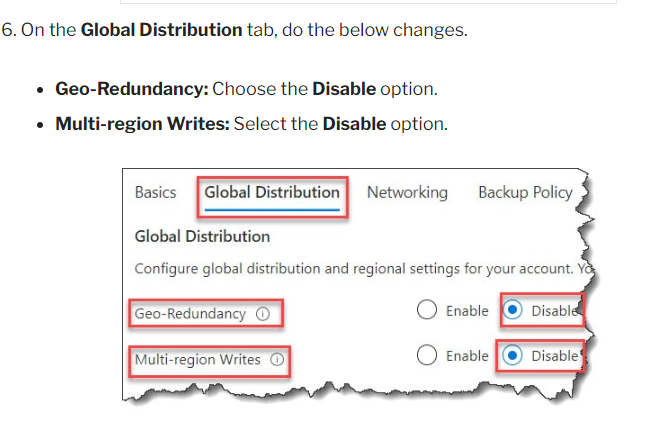
* [**Subscription**](https://azurelessons.com/you-are-not-eligible-for-an-azure-subscription/)**:** Select the Azure Subscription that you want to use here to create the Azure Cosmos DB Account.
* [**Resource Group**](https://azurelessons.com/what-is-resource-group-in-azure/)**:** Select the existing resource group or if you don’t have an existing resource group, you can click on the Create new link to create a new resource group.
* **Account Name:** Provide a unique name for the Azure Cosmos DB Account.
* **Location**: Select the region or location.
* **Capacity mode**: Choose the capacity mode.
* **Apply Free Tier Discount:** Choose the Apply free tier discount option based on your need.



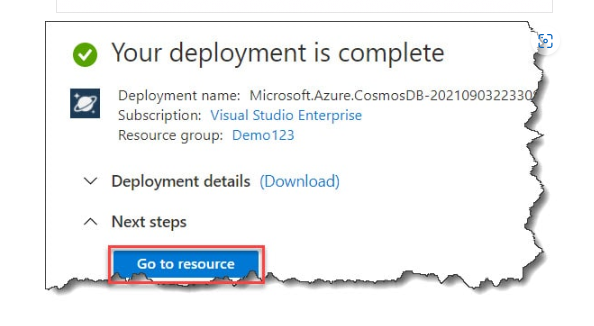




Provisioned throughput which will give you a consistent performance throughout or you can go for the serverless, which gives you the outer scale abilities mostly choose throughput option.



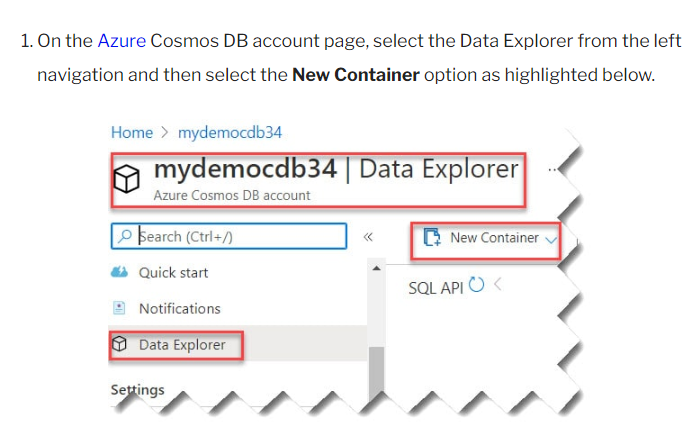
Keep the other tab options as the default value as it is and then click on the **Review + Create** button.

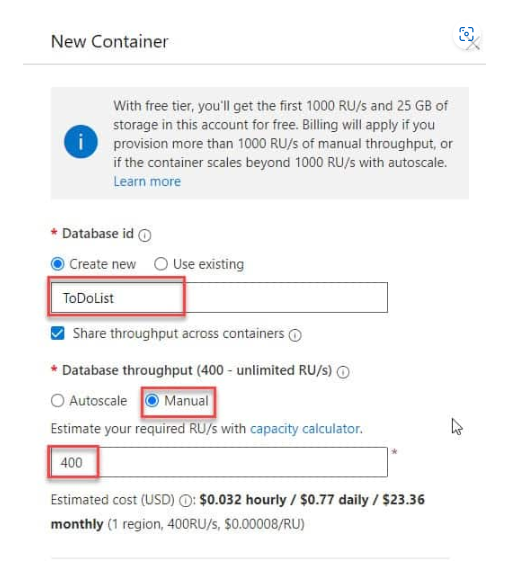


After complete go into Feature and select Azure synapse link and turn on them

Now we create cosmos DB account let’s create the database and the container

Click on the dropdown of the New Container choose Database

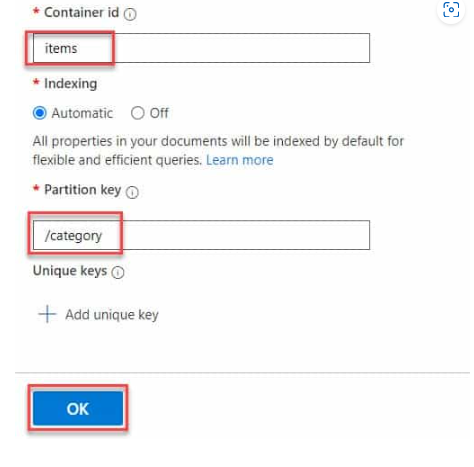




By clicking on the manual, you can specify what the RU you want to use for our project so ex: - if it’s 400 you just specify that value and that would be provisioned throughout but you can auto scale by adding auto scale, you can save some cost ex: - if you specify that you are max RU is 1000, it’s going to start from 100 RU which is 10% of the maximum value and then it be scaled up to 1000 all year depending on your workload

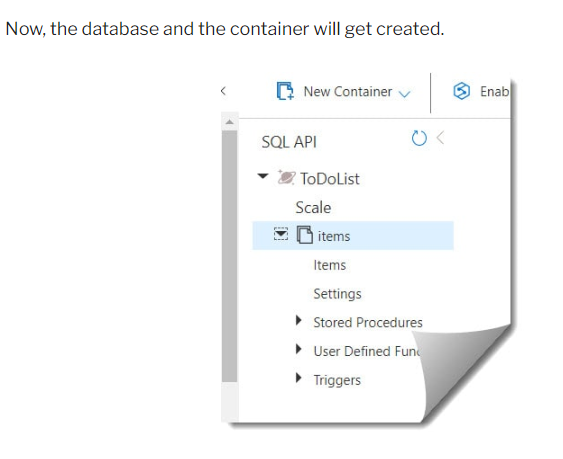
Click on OK

your database is created. click on three dots. click on new container

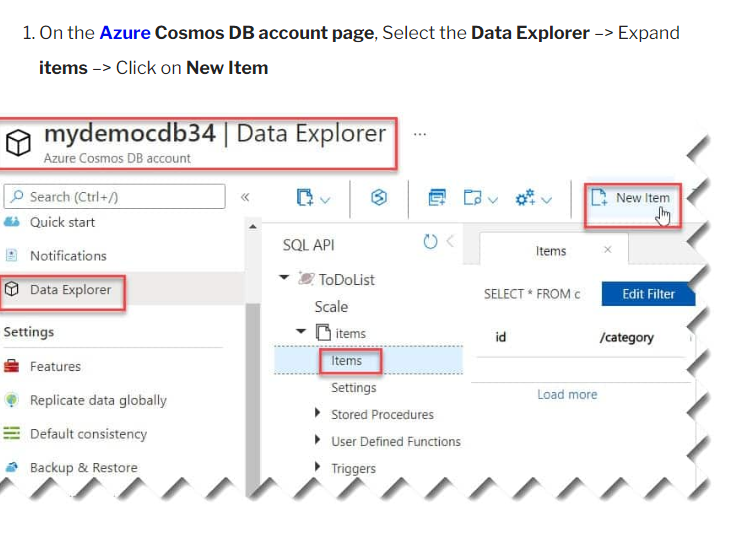


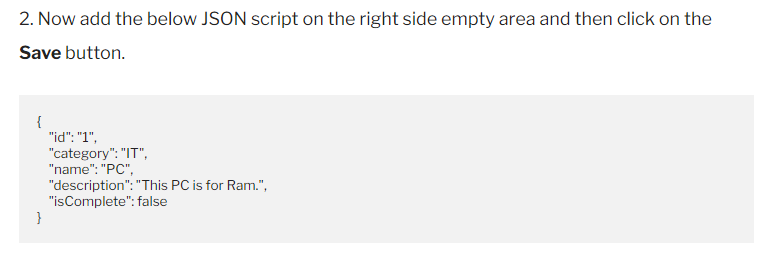
Partition key = Partition key is a key which we got a high cardinality so the Cosmos DB can distribute the data and distribute the workload as well

There one section also which is Analytical store which you can make ON or OFF, so this gives us the ability to push the data from a transactional store to the analytical store, replicate the data from transactional store to the analytical store if you leave that OFF, then nothing will be replicated.



Here for example ToDoList is a database and items are a container Here you will see items which list all the items within the container, settings let you set your time to leave information for the transaction store as well as an analytical store.





(Upper Wala khali example ha)

When you save it look like this



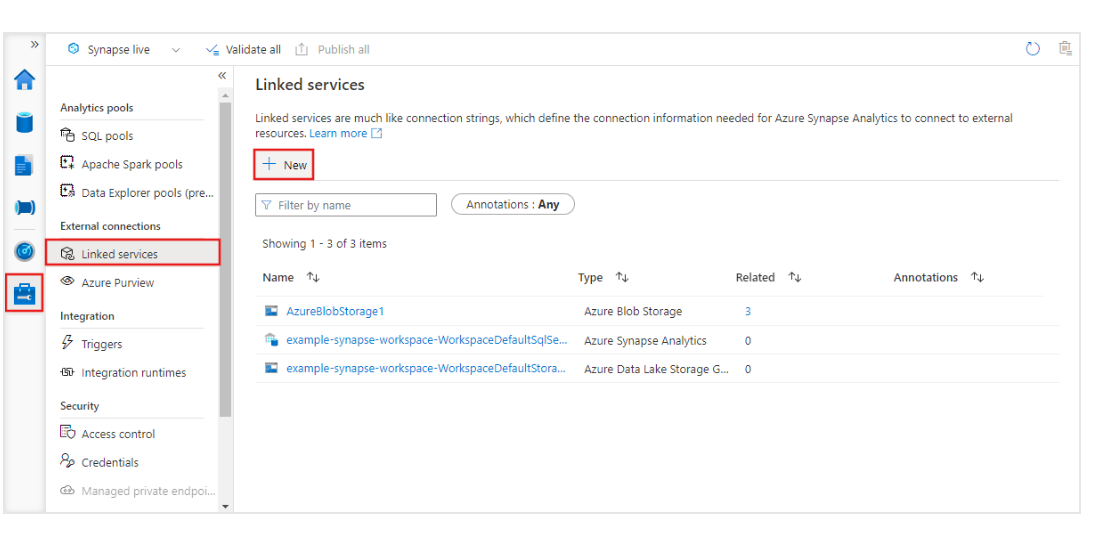
App jub save karo ga tho app ko ek query deka ga jo ke ha SELECT\* FROM c jiska meaning ha select all data from container

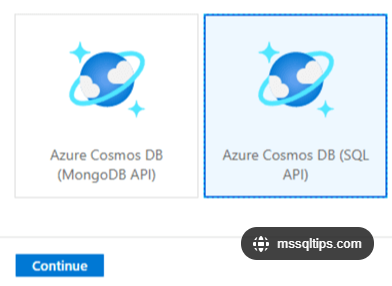
This data is going into the transactional store. That data is not in the analytical store

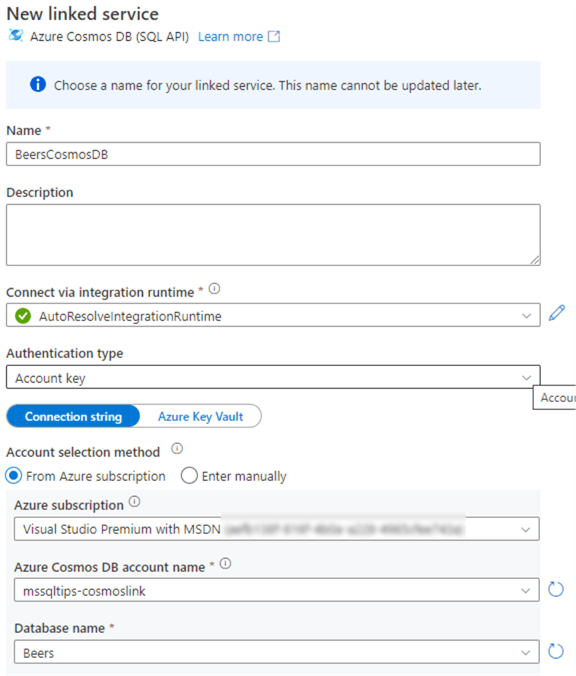
**If you want to save another entry, then select the New Document again and save the entry with another unique id and change the other properties accordingly.**

[How To Create An Azure Cosmos DB Account - Azure Lessons](https://azurelessons.com/create-azure-cosmos-db/)

In this we are going to create link service within synapse studio and connect to data in our cosmos DB analytical store and query that data as well

On a linked service you must search **Cosmos**





After that test the connection and click on OK

Steps: -

1. Go the Data hub in that there is a Linked section in that your Azure cosmos DB is created under that your linked service is present and under that your container is there click on three dots
2. After click on three dots select New SQL script, select Top 100 rows
3. We need to create a credential to access the cosmos DB and we need to specify the secret
4. To get the secret go to cosmos DB in the left side there is a **key** click them and go to Read-only keys

So, what we done is using synapse serverless SQL and connect to **Built-in pool** here and successfully quired the data from cosmos DB

One thing here is that we haven’t specified the analytical store anywhere and we didn't say that we access data from analytical store but serverless SQL pool only lets you access the data from your analytical store. So, whenever you are running queries from serverless SQL pool it’s always accessing the analytics store not from transactional store.

Using Spark pool

Steps: - 1) Click on the three dots of the container click on the new notebook after clicking you get different options such as Load to Data Frame, write DataFrame to Container, Create Spark Table and etc.

write DataFrame to Container = The spark pool gives the ability to write the data back into Azure Cosmos DB container ex: - your use case might be that we want to read data from your analytical store do some ML task and write the result back into Cosmos DB Transactional store so that’s used by the business applications (cosmos.oltp)

Load to DataFrame = Here we use (cosmos.olap) which is analytical processing engine to get the data, so this is going to read the data from your analytical store rather than the transactional store and that we do for analytical process

2) we must choose Load to DataFrame

Power BI Integration

Connecting from PowerBI Desktop

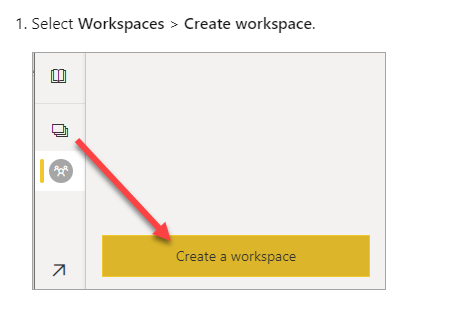
In this we are going to use power BI desktop application to access the data in our serverless SQL pool and create reports on them.

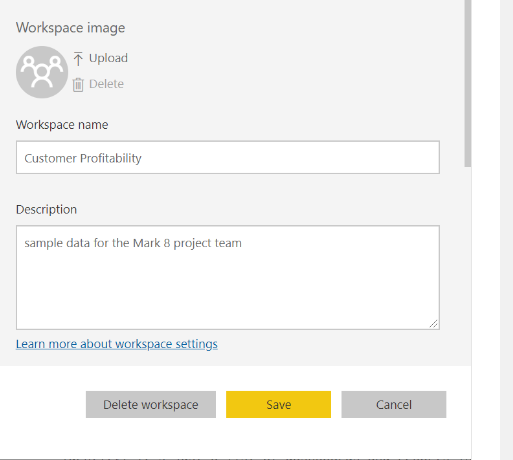
1. Open the Power BI desktop and click on the **Get data**
2. You will get different option to get the data from that choose **Azure**
3. When you choose the azure, you will see different option in that choose **Azure synapse analytics SQL** and click on connect
4. After click on connect you will ask server and database, Data Connectivity mode which is Import and Direct Query
5. And for server details go to Managed hub under that synapse studio and in that go the SQL pool in that there is a Built-in click on them you will see the **workspace SQL endpoint** copy that and place over the server details and you can choose Import or DirectQuery and from them **choose Directquery**
6. Now we need to use credentials to authentication for that choose to use your Microsoft account
7. After signing we can see different databases and select one table and apply visualizations.

Publish to power BI workspace

Agar app ka pass work mail id raha ga tho app ko ya sub create karna ke koi jurat nahi ha.

* If you use the mail that has azure subscription it won’t work. So, you must create a new mail id for that go Azure AD, click on + add (User) and give username and password also.
* We don’t want to be putting things in my workspace because that can’t be shared
* Jub app ya sub kar lan ga tho app kay karan ga ke workspace create karan ga lakin who app ko bola ga ke upgrade your power BI thub app upgrade kar ka phir sa workspace creates karan ga (When you go again to create a workspace It asks to Upgrade to Power BI pro, again go in the workspace after the upgrade)





1. After you give workspace name
2. Go to the power BI desktop and sign in and for that use the mail id and password which you use during the creation of power BI workspace
3. After sign in click on Publish All when you click on Publish All it asks workspace name

If you click the report, now you notice that the report is not working (means it’s not shown any thing) and it tells that there is missing credential in dataset an to fix that, you can go on the main screen of your workspace where you see your report and dataset, click on the three dots over the dataset and click on setting and it’s tell you that your data source can’t be refreshed because credentials are invalid this is because the user we created doesn’t have access to any of the resource that we have created in azure portal which is things like storage account, and the synapse studio things like that so we need to fix that.

Access to power BI

What we done

1. We create new Azure Activity Directory User
2. We use that user to create new power BI user and update in power BI pro
3. We create new workspace

What we have do now to fix the issue

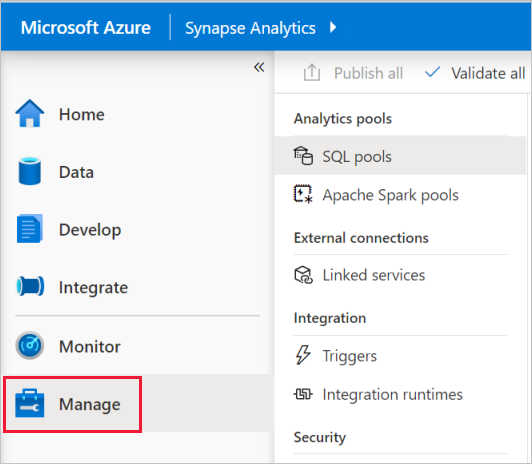
1. We can give the user access to the synapse resource group that we created for the project this will give the user access not only the resource group, but also all the resource in that resource group
2. We still need to provide access separately to the synapse studio
3. We need to provide a specific role called storage blob data contributor to user so that the user can read or write data to the storage account

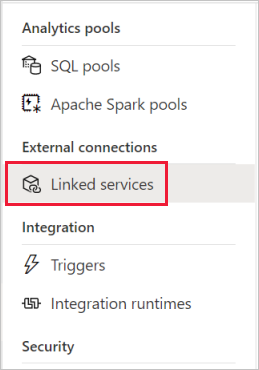
Steps: -

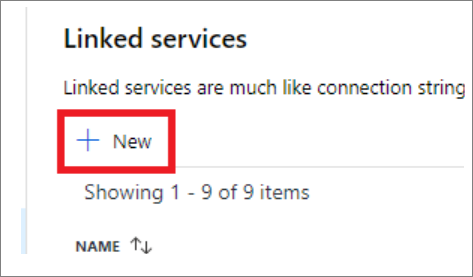
1. Login to the main user account
2. Go to Resource group choose the resource group that have storage, synapse, cluster pool
3. Click on IAM under that click on + add under that choose **add role assignment** under that choose **contributor** after that choose select member and, in that search, your mail id which you create through Azure AD and after select click on Review+assign
4. For provide the access to synapse studio separately: Go to your resource group where your storage account, synapse, cluster pool
5. Click on synapse and click on **open synapse studio** and sign with main account go to the manage and click on the access control and we need to add new user so click on +add and select the **synapse administrator** and select user which is the user you created through Azure AD (not main user) click apply
6. Last, we want to provide this user, access to the storage account: Go to resource group again chose the storage account after that go to the IAM (access control) click on +Add under that chose **Add role assignment**
7. Search **storage blob data contributor** and click next and select the member (users) which you created through Azure AD and click Review+assign

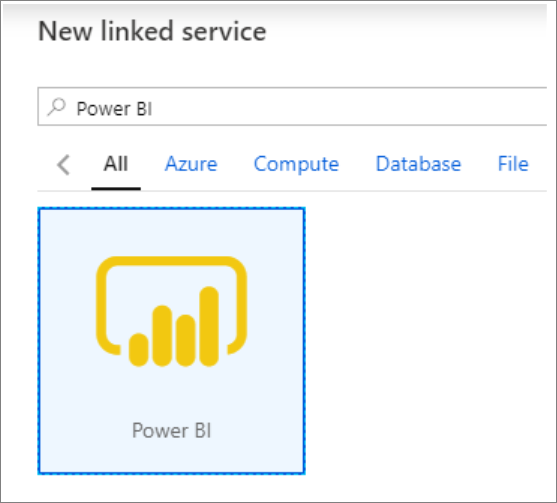
Synapse studio power BI integration

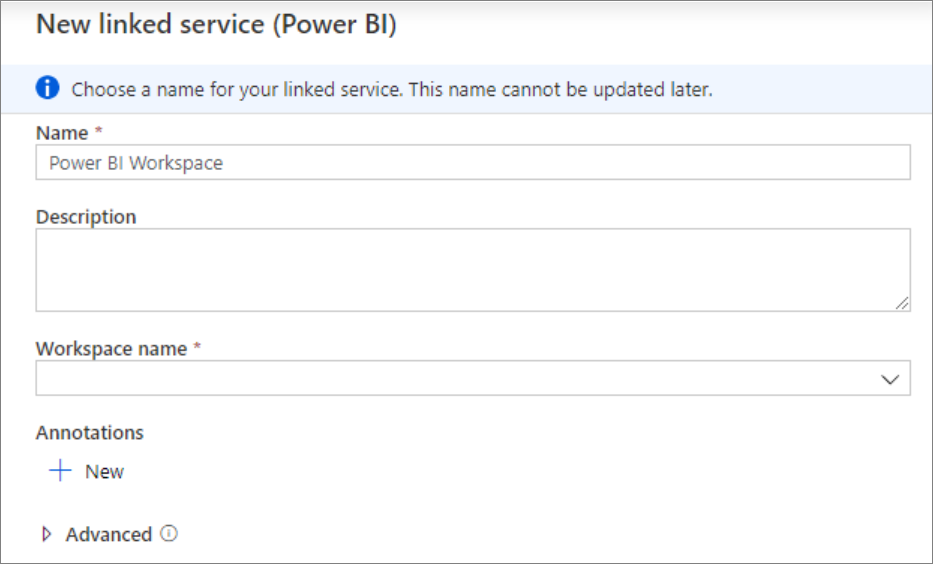
## Link Power BI workspace to your Synapse workspace



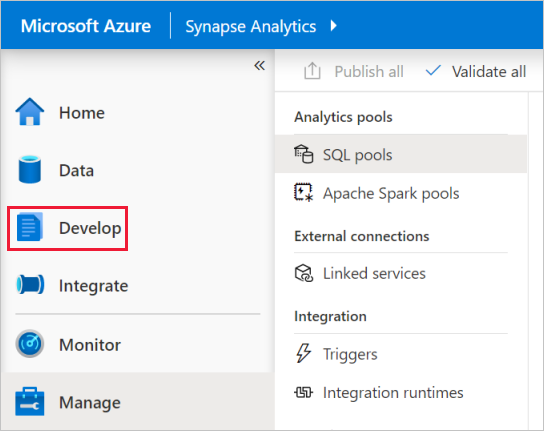




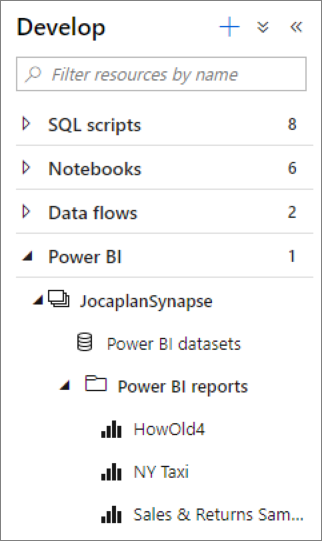




Click on created



Now when you go to your develop hub, you are going to be able to develop reports on the power BI workspace

 Here JocaplanSynapse is your linked service

When you click on power BI dataset under your linked service, you will notice that there is no modelling tool available, so you can’t do any kind of data modeling and there is no ability to add dataset for suppose you click on **+ New Power BI dataset** it is going to give us a file extension of .pbids and that will open up in the power BI desktop where you can model and then came back to develop your reports, but in terms of reports when you click on that you are going to have entire capability that you get within the power BI desktop

So, the takeaway here is that we’ve got a full integration of the power BI, report developer tool but not the datasets and if we are doing any scheduling of datasets or anything like that, we must come into the power BI workspace to do that.