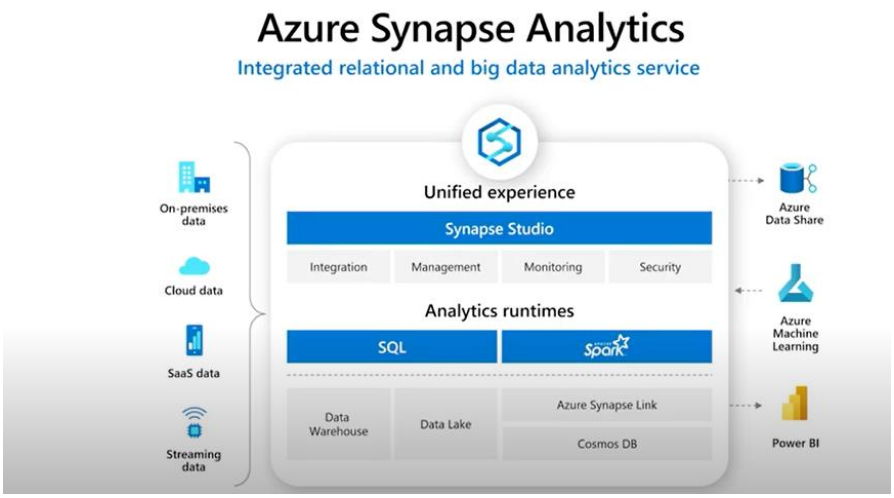
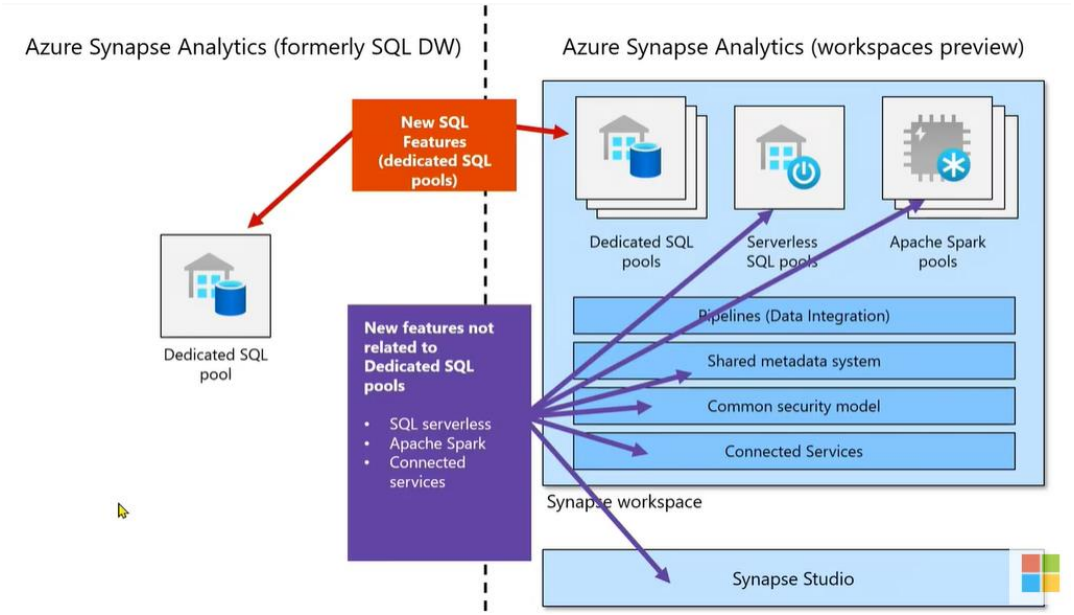
Introduction to Azure Synapse Analytics

What is Azure Synapse Analytics?



- Azure Synapse Analytics brings together **data warehouse** and **big data analytics** and **Data integration** into a single and unified space workspace.
- It allows customers to build end-to-end analytics solutions and perform data ingestion, data exploration, data warehousing, big data analytics, and machine learning tasks from a single, unified environment.
- The advantage of having a single integrated data service is that, for enterprises, it accelerates the delivery of BI, AI, machine learning, Internet of Things, and intelligent applications and Data professionals of all types can collaborate, manage, and analyze their most important data efficiently—all within the same service
- Azure Synapse Analytics is deeply integrated with Power BI and Azure Machine Learning to greatly expand the discovery of insights from all your data and apply machine learning models to all your intelligent apps.
- It offers Synapse SQL Engine, Apache Spark Engine and Data Integration engine.
- It provides deep integration of Apace spark and SQL Engine.
- **Synapse SQL** is a distributed query system for T-SQL and offers **serverless** and **dedicated** resource models
- **Apache Spark for Azure Synapse is** used for data preparation, data engineering, ETL, and machine learning.
- **Data Integration engine** provides experiences as Azure Data Factory, allowing you to create rich at-scale ETL pipelines without leaving Azure Synapse Analytics.
- It provides Unified management, monitoring, and security.

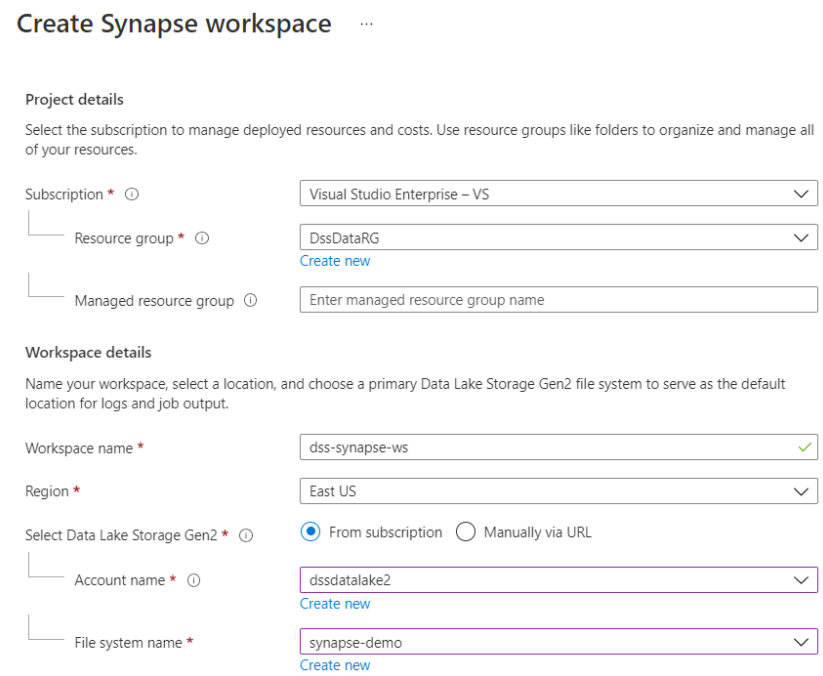
Azure Synapse Analytics workspace



- A workspace is the top-level resource and comprises your analytics solution
- Synapse SQL offers both **serverless** and **dedicated** resource models. Both supports Data Warehousing and Data Lake
- It has one default serverless SQL Pool which maps to distributed query service.
- There can be any number of dedicated SQL Pools and any number of Apache Spark Pools
- Pipeline Provides Data integration, Orchestration and Data Movement.
- Shared metadata system makes it easy to share tabular data between SQL and Spark.
- Entire workspace, all resources, all pools are governed by common security model, which makes it easy to manage.
- There are series of connected services which expands the reach of synapse in other services.
- Synapse Studio is one stop shop for data engineers to code, monitor, manage, debug, secure.

Lab 1: Create Synapse Analytics Workspace

Search→synapse→Azure synapse Analytics→Create→



Security Tab→Provide password for administrator access to the workspace's SQL pools.

Dedicated SQL Pool:

- Dedicated SQL pool (formerly SQL DW) represents a collection of analytic resources that are provisioned when using Synapse SQL.
- The size of a dedicated SQL pool is determined by Data Warehousing Units (DWU).
- Dedicated SQL pool uses PolyBase to query the big data stores. PolyBase uses standard T-SQL queries to bring the data into dedicated SQL pool (formerly SQL DW) tables.
- Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage.

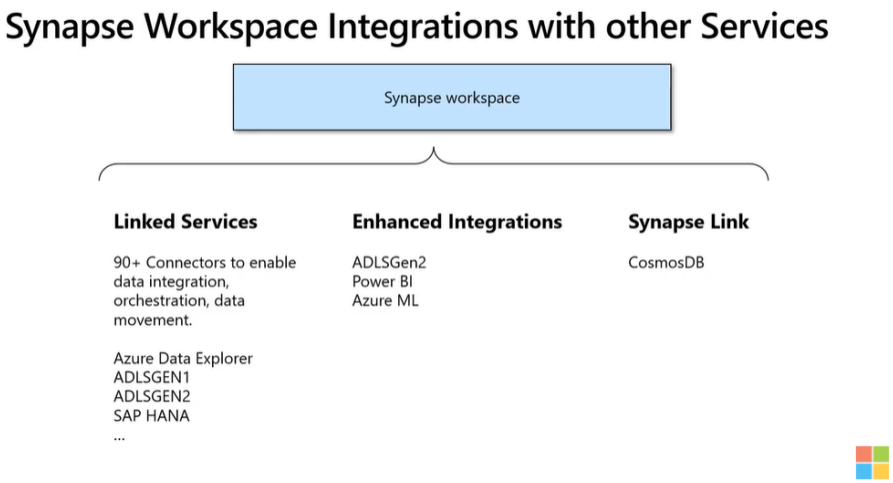
Serverless SQL Pool:

- Serverless SQL pool is a query service over the data in your data lake.
- Serverless SQL pool is a distributed data processing system, built for large-scale data and computational functions. Serverless SQL pool enables you to analyze your Big Data in seconds to minutes, depending on the workload.
- Serverless SQL pool is serverless, hence there's no infrastructure to setup or clusters to maintain.
- There is no charge for resources reserved, you are only being charged for the data processed by queries you run, hence this model is a true pay-per-use model.
- You can use following tools for querying Data: Azure Synapse Studio, Azure Data Studio ,SSMS

Spark Pool:

- Spark pools in Azure Synapse offer a fully managed Spark service.
- Apache Spark provides primitives for in-memory cluster computing.

- Apache Spark is a parallel processing framework that supports in-memory processing to boost the performance of big-data analytic applications.
- Apache Spark includes many language features to support preparation and processing of large volumes of data so that it can be made more valuable and then consumed by other services within Azure Synapse Analytics



Synapse Studio:

- Synapse Studio features a user-friendly, web-based interface that provides an integrated workspace and development experience.
- This allows data engineers to build end-to-end analytics solutions (ingest, explore, prepare, orchestrate, visualize) by performing everything they need within a single environment

Data:

- You can create Database and Database objects
- You can create linked Database to access data from external Repositories.
- By default, the Azure Data Lake Storage Gen2 account, which is provided during the creation of the Synapse workspace is linked and shown here.
- Based on the repository, different options can be seen on the toolbar like creating a new SQL script, new notebook, new data flow, new dataset, as well as file-based operations like creating or deleting a new file or folder

Develop:

- It provides options to create new artifacts like SQL script, Notebook, Data flow, etc.

Integrate:

- We can create data pipelines, jump directly to the Copy tool which allows us to create data pipelines step by step using a wizard, or browse a gallery of samples or previously created data pipelines to reuse the same for integrating data.

- Monitor:
- Synapse Studio is not only a developer console but also an administrative console as well. With Monitor view you can monitor the pipeline executions, triggers that initiated a pipeline execution, and different integration runtimes.
- It also provides different options to monitor spark applications and those job executions that are generated from those applications, ad-hoc SQL queries or requests that are executed, as well as options to debug a data flow as well.

Manage:

- In Analytics pools section ,You can see built in serverless pool and create new SQL Pool.
- One can create linked services to register external data repositories in the external connections section.
- In the Integration section triggers and Integration runtime can be registered.
- In the Security section, one can configure access control to this environment to different users and group, modify the credentials that we configured for administrative access, and manage any private endpoints for secure network connectivity (if any).

Launch Synapse Studio:

Option 1:Workspace→Open Synapse Studio→Open

Option 2:https://web.azuresynapse.net/

Using SQL Pool

Lab2: Create Dedicated SQL Pool:

1. Launch Synapse Studio: Workspace→Open Studio OR https://web.azuresynapse.net/
2. Manage→SQL Pools→New→Specify Name and Performance Level→Create

Knowledge Center:

- Knowledge Center accelerate developer learning how to use synapse by providing sample SQL Scripts, notebooks, Pipeline Templates and easy access to data from Azure Open Data.
  Home→Learn

Lab3: Load New York Taxicab Data From blob storage to Data Warehouse Using SQL Script.

1. Home Hub→Learn→Browse Gallery→ Select SQL Script Tab →Select Load the New York Taxicab Dataset→ →Continue→
2. **SQL pool→Select an existing pool** → Select **SQLPOOL1**→ and select the **SQLPOOL1** database → **Open Script**.

3. You can make required changes to script and run it.

Lab 4: Link Sample Dataset

1. Home Hub→Learn→Browse Gallery→Select Dataset Tab→Select Any Dataset→Continue→Add Dataset
2. Data Hub→Sample Dataset→Observe the newly added Dataset→Right Click→New SQL Script→Select TOP 100 rows

Note: You can also use New Notebook option and work with it using Spark Cluster

Lab 5: Link GreenTaxidataset present in datalake in Synapse

Data→Linked→Add new resource→Connect to external data→Data Lake Gen2

Lab 6: Explore Sample datasets with Serverless SQL Pool

1. Home Hub→Learn→Use Samples Immediately→Query Data With SQL

Copy Statement:

The COPY statement is the most flexible and secure way of bulk loading data in Synapse SQL.

Refer: COPY INTO (Transact-SQL) - (Azure Synapse Analytics) - SQL Server | Microsoft Docs

Authentication mechanisms with the COPY statement - Azure Synapse Analytics | Microsoft Docs

Example:

COPY INTO dbo.[lineitem] FROM

'https://unsecureaccount.blob.core.windows.net/customerdatasets/folder1/lineitem.csv'

The COPY statement's defaults match the format of the line item csv file.

Note: You can now skip header rows for delimited text files on Azure SQL DW by using the First_Row option in the external file format.

Lab 7: Create table for yellow taxi data in the Pool.Use Copy command to bulk load data from datalake to sqlpool .

```
Create Table YellowTaxiTrip
(
VendorID int,
tpep_pickup_datetime     datetime,
tpep_dropoff_datetime    datetime,
passenger_count int,
trip_distance      float,
RatecodeID         int,
```

```
store_and_fwd_flag          varchar(2) COLLATE SQL_Latin1_General_CP1_CI_AS NOT NULL,
PULocationID       int,
DOLocationID       int,
payment_type int,
fare_amount        money,
extra      money,
mta_tax_tip_amount         money,
tolls_amount       money,
improvement_surcharge   money,
total_amount  money
)
WITH
(
   DISTRIBUTION = ROUND_ROBIN,
   CLUSTERED COLUMNSTORE INDEX
);

COPY INTO dbo.YellowTaxiTrip
FROM 'https://dssdatalake2.dfs.core.windows.net/taxidata/YellowTaxiTripData_201812.csv'
WITH (
   FILE_TYPE = 'CSV',
   CREDENTIAl= (IDENTITY='Shared Access Signature', SECRET='sp=r&st=2021-04-04T09:08:13Z&se=2021-04-
04T17:08:13Z&spr=https&sv=2020-02-10&sr=b&sig=jIDmjdyHTiaApifpkIysqbCZjEfWMzcfaIq577DKbLk%3D'),
    ROWTERMINATOR='\n',
    FIRSTROW = 2
)
```

Lab 8: Ingest NYC Taxidata into Dedicated Pool

1. Develop→New SQL Script→Use “LoadNycTaxidata” script

Lab 9: Explore Taxidata

1. Data Hub→Databases--> Select Your Dedicated pool→Expand Tables→TaxiTrip→New SQL Script→Select Top 100 Rows
2. Get total trip distance and average trip distance based on Passanger Count

```
SELECT PassengerCount,
SUM(TripDistanceMiles) as TotalTripDistance,
```

```
    AVG(TripDistanceMiles) as AverageTripDistance
    FROM dbo.TaxiTrip
    WHERE PassengerCount > 0 AND TripDistanceMiles > 0
    GROUP BY PassengerCount
ORDER BY PassengerCount
```

Note: You can quickly change the view to Chart to see a visualization of the results as a line chart

Enabling Synapse workspace features on an existing dedicated SQL pool:

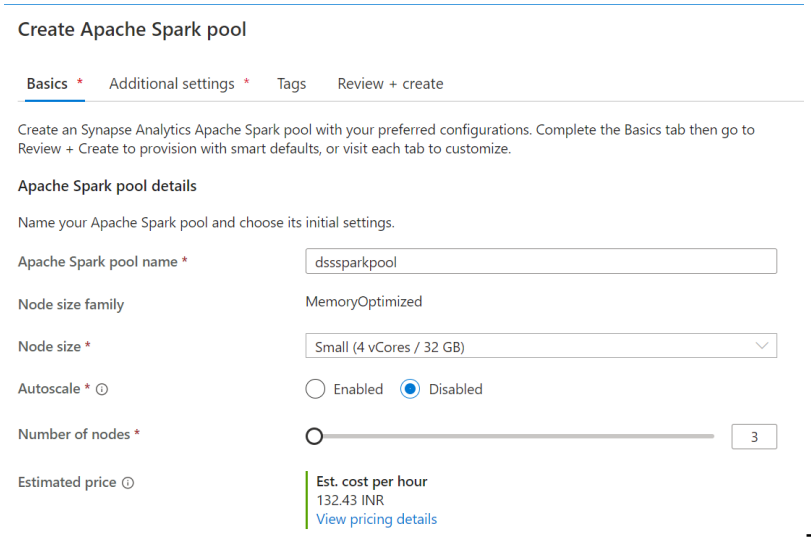
- Select existing Azure SQL DW→Overview→New Synapse Workspace
- This new capability will allow you to connect the logical server that hosts your existing data warehouse instances to a new Synapse workspace.
- All the data warehouses hosted on that server are made accessible from the Workspace and Studio and can be used in conjunction with the Synapse partner services (serverless SQL pool, Apache Spark pool, and ADF)

Refer: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-elt-data-loading#what-is-elt

Working With Spark Pool

Lab 10: Create Spark Pool

1. Manage→Analytic Pool→Apache Spark Pool→



→Review+Create.

Lab 11: Explore the Linked Taxi Dataset

1. Data Hub→Linked→Sample Dataset→Select Dataset→Right Click→New Notebook→Load To DataFrame
2. Attach Notebook to SparkPool→Run all

Lab 12: Ingesting SQL pool data into a Spark database and Analyse it with Notebook

1. Develop→Add New Resource—Notebook→Add Following Code

```
%%spark
spark.sql("CREATE DATABASE IF NOT EXISTS sparknyc")
val df = spark.read.sqlanalytics("learningpool1.dbo.TaxiTrip1")
df.write.mode("overwrite").saveAsTable("sparknyc.taxitrip")
Note: In Data Hub Observe the Database and the Table created in Spark.
```

Lab 13: Analyse Data using Spark and Notebook

```
df = spark.sql("""SELECT PassengerCount,
SUM(TripDistanceMiles) as TotalTripDistance,
AVG(TripDistanceMiles) as AverageTripDistance
FROM sparknyc.taxitrip
WHERE PassengerCount > 0 AND TripDistanceMiles > 0
GROUP BY PassengerCount
ORDER BY PassengerCount""")
display(df)
df.write.saveAsTable("sparknyc.passengerstats")
```

Lab 14: Ingesting Spark table data into an SQL pool table

```
%%spark
val df = spark.sql("SELECT * FROM sparknyc.passengerstats")
df.write.sqlanalytics("sqlpool001.dbo.PassengerStats",Constants.INTERNAL)
```

Integrate With Pipeline

- The **Integrate** hub allows you to build data pipelines and perform code-free data transformations.

- In Azure Synapse Analytics, the data integration capabilities such as Synapse pipelines and data flows are based upon those of Azure Data Factory.
- Refer to following link to know the difference between Integration in Synapse Analytics and Data Factory

https://docs.microsoft.com/en-us/azure/synapse-analytics/data-integration/concepts-data-factory-differences

Lab13: Use Copy Activity to Copy file from Data Lake to Blob

Lab 14: Load Data From table in Azure SQL to SQL Pool

Lab: Ingest data into Dedicated SQL Pool

Lab 15: Transform Data Using Mapping Data Flow

Use MoviesDb.csv

1. Filter movies of genre comedy that came out between the years 1910 and 2000

   Filter Transformation :toInteger(year) >= 1910 && toInteger(year) <= 2000 && rlike(genres, 'Comedy')

2. Aggregate the Data based on average rating of comedy movies by year

   Group by :year

   AvgRating: avg(toInteger(Rating))

3. Load the Final Dataset to Data Lake

Note: To learn Data Flow Expression language

https://docs.microsoft.com/en-us/azure/data-factory/data-flow-expression-functions?toc=/azure/synapse-analytics/toc.json&bc=/azure/synapse-analytics/breadcrumb/toc.json

Monitor Hub

Monitor hubs Provides you a history of all the activities taking place in the workspace and which ones are active now.

- Under **Integration**, you can monitor pipelines, triggers, and integration runtimes.
- Under **Activities**, you can monitor Spark and SQL activities.