eBook series

Microsoft Azure

Limitless Analytics with Azure Synapse

Limitless Analytics with Azure Synapse

Introduction to Azure Synapse

Azure Synapse Analytics, formerly known as Azure SQL Data Warehouse, is not a mere data warehouse any more; it is an amalgamation of big data analytics with an enterprise data warehouse. Azure Synapse is a limitless analytics platform that provides a unified experience to discover and explore data quickly to find meaningful insights at scale. You have the option to select an analytics runtime as per your business requirements. Azure Synapse not only makes it easy to start and scale in the cloud, but you also save costs. *Figure 1* displays all the components of Azure Synapse tied together within a unified experience called Azure Synapse Studio. We will learn more about all these components later in this book.

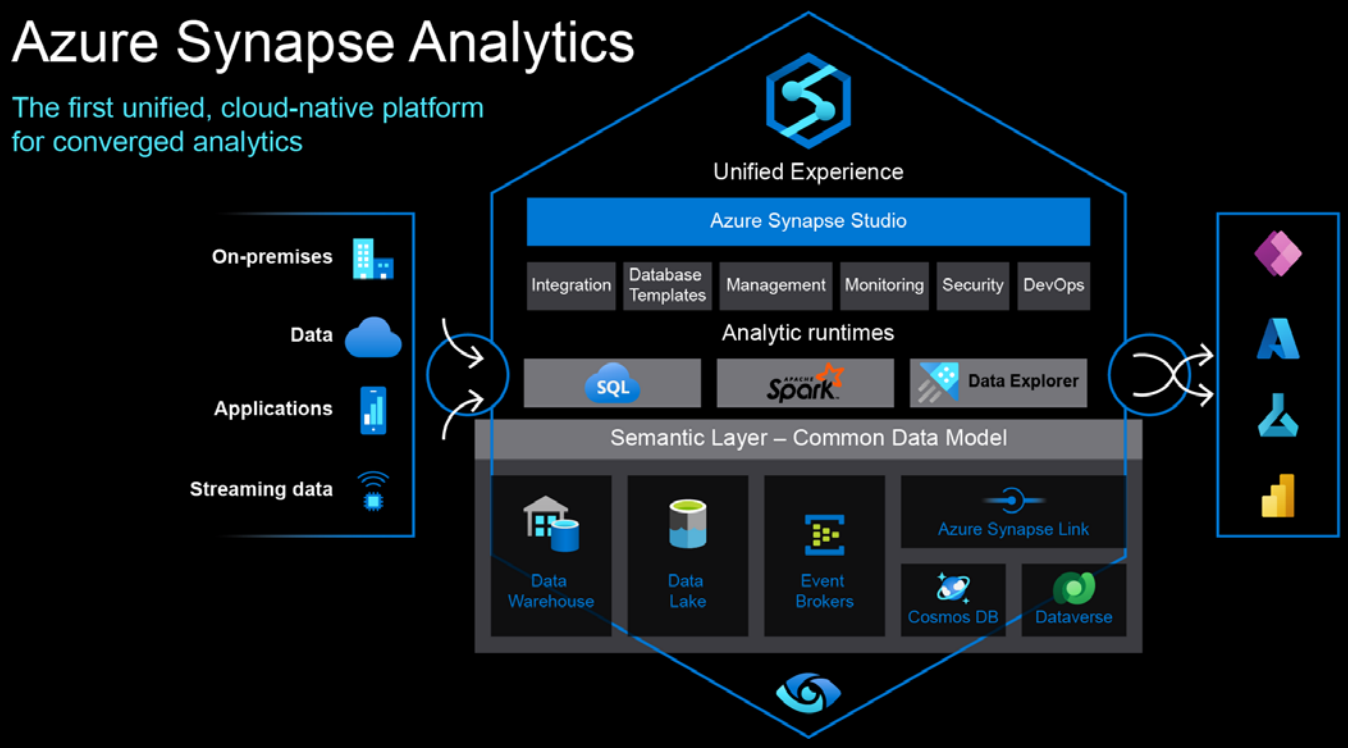

Figure 1: Different components of Azure Synapse Analytics

Before we dive deeper into the various features offered by Azure Synapse, let's try to learn more about the Synapse workspace and Synapse Studio in the following section.

1

Getting started with your first Azure Synapse project

In order to get started with Azure Synapse, we need to create a Synapse workspace. A Synapse workspace provides an integrated console to manage, monitor and administer all the components and services of Azure Synapse.

Refer to Quick-start: Create a Synapse workspace to create your first Synapse workspace.

You can connect to your workspace using **Synapse Studio**. Synapse Studio is a free web tool provided by Azure Synapse for all data engineers, data scientists and report developers. Synapse Studio also enables you to manage and monitor all your resources created under your Synapse workspace.

If you are new to Azure Synapse, it is highly recommended that you explore all the resources available under **Knowledge centre** in Synapse Studio. Go to **`Browse gallery`** in the Knowledge centre and go through the available templates, datasets, notebooks, SQL scripts and pipelines to get yourself acquainted with Azure Synapse. To learn more about the Knowledge centre, you can go through the Explore the Synapse Knowledge centre documentation.

Let's explore the Integrate hub, which is used to ingest and orchestrate data, before we proceed further with other capabilities.

2

Data engineering

Azure Synapse allows you to create new data pipelines to perform one-time or scheduled data ingestion from more than a hundred different data sources.

The Integrate hub gives you multiple options to bring your data to Azure Synapse and to orchestrate your data pipelines. The **Copy Data** tool is a code-free integration tool that can be used to copy data from a source to Synapse.

You also have the option to use the code-first integration tool to add one or many transformations to your data by creating pipelines. A Synapse pipeline can be created by using one or more activities, which can be connected to each other by dependency endpoints. By default, you get a **`Success`** endpoint, but you can change this to **`Failure`**, **`Completion`** or **`Skipped`** if required, as you can see in *Figure 2*:

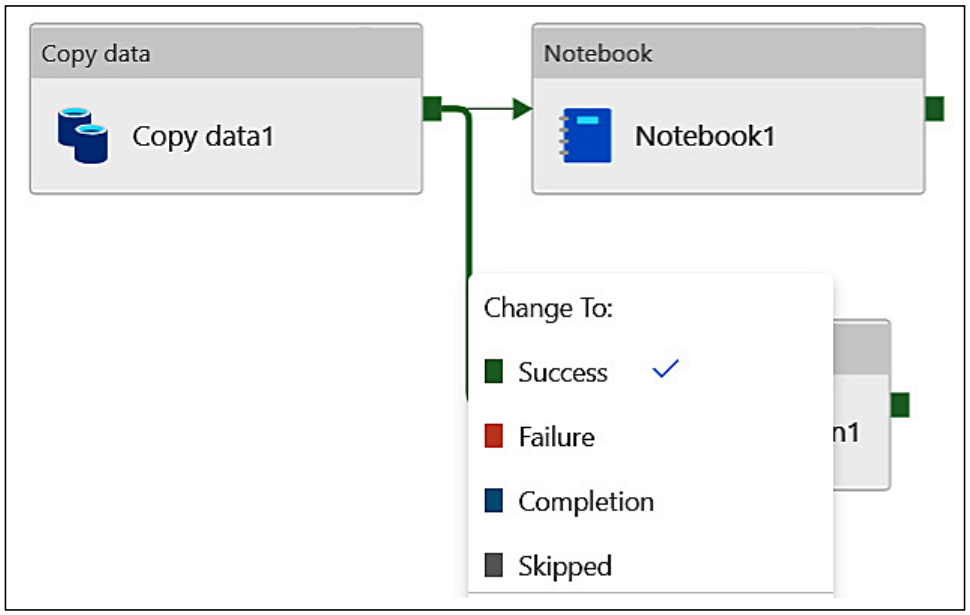

Figure 2: Creating dependency endpoints for two activities in a Synapse pipeline

You can also loop through any activity by moving that activity inside an iteration activity such as the **ForEach** or **Until** activities.

To learn more about Synapse pipelines and natively supported transformation activities, read through the [Pipelines and activities in Azure Data Factory and Azure Synapse Analytics] documentation.

There are many templates already available within the Knowledge centre that can be leveraged to create pipelines instead of creating them from scratch. You can click on **`Import from pipeline template`** from the drop-down list and start exploring the templates.

In the following section, *Big data*, we are going to learn how to use Azure Synapse for big data.

3

Big data

Now you can create and optimise Apache Spark pools on Azure Synapse with autoscaling and query optimisation features.

Apache Spark is a very fast unified analytics engine for **big data** and **machine learning**. Synapse Spark pool is one of Microsoft's implementations of **Apache Spark** in Azure. Synapse Analytics workspace has a Spark engine built in, along with notebook support. Because Synapse Spark supports C#, we can write Spark .NET directly within notebooks. You can also write your code in **Python**, **Scala**, **C#** and **SQL**.

One Spark pool can be accessed by multiple users, but for every user, one new Spark instance will be created. A Spark instance is also dependent on the Spark pool capacity: if there is enough capacity in the pool to run multiple queries, the existing instance will be able to process the job; otherwise, a new instance will be created to process the job.

Figure 3 displays different components of Apache Spark on Azure Synapse:

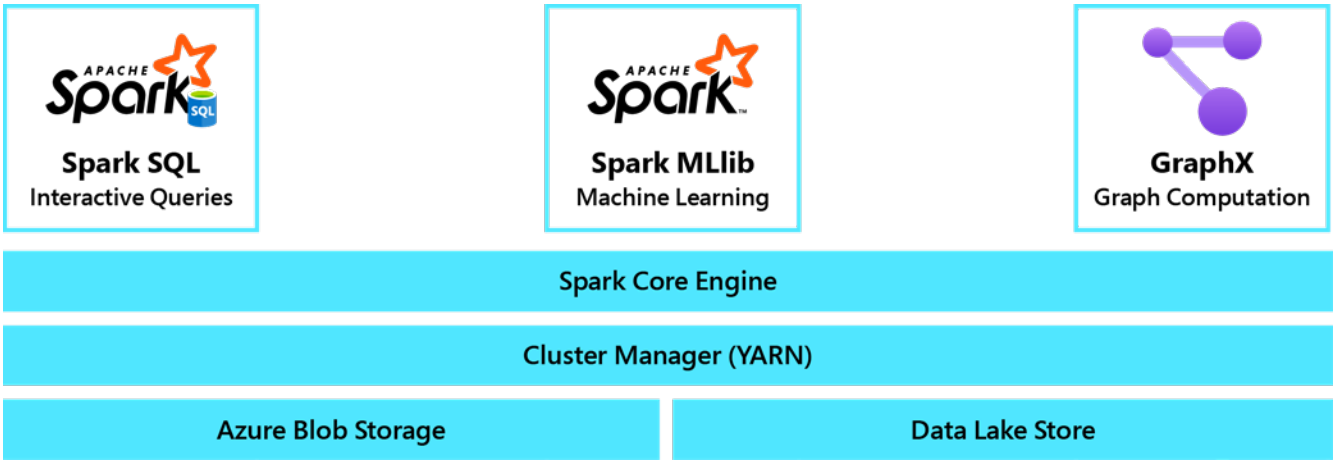

Figure 3: Apache Spark in Azure Synapse

Go through Quick-start: Create a serverless Apache Spark pool using Synapse Studio to create your first Spark pool in Azure Synapse.

Serverless SQL is one of the best features of Azure Synapse, which can be used to explore data directly from various sources without copying it to any relational database and we are going to learn more about this feature in the following section.

4

Serverless data lake exploration

Serverless SQL pool is a serverless distributed data processing system that enables you to analyse your big data faster. You don't need to provision any compute or maintain the scalability. In Serverless SQL compute, scaling automatically accommodates the resource requirements for any query. The Serverless SQL architecture also has a control node and compute nodes, but it does not have a **Massively Parallel Processing** (**MPP**) engine; instead, it uses a **Distributed Query Processing** (**DQP**) engine.

The architecture, as illustrated in *Figure 4*, explains how a **control node** leverages a DQP engine to distribute a query across various computes as per requirements. **Compute nodes** will reach out to the storage to fetch the required data as requested and send it back to the control node.

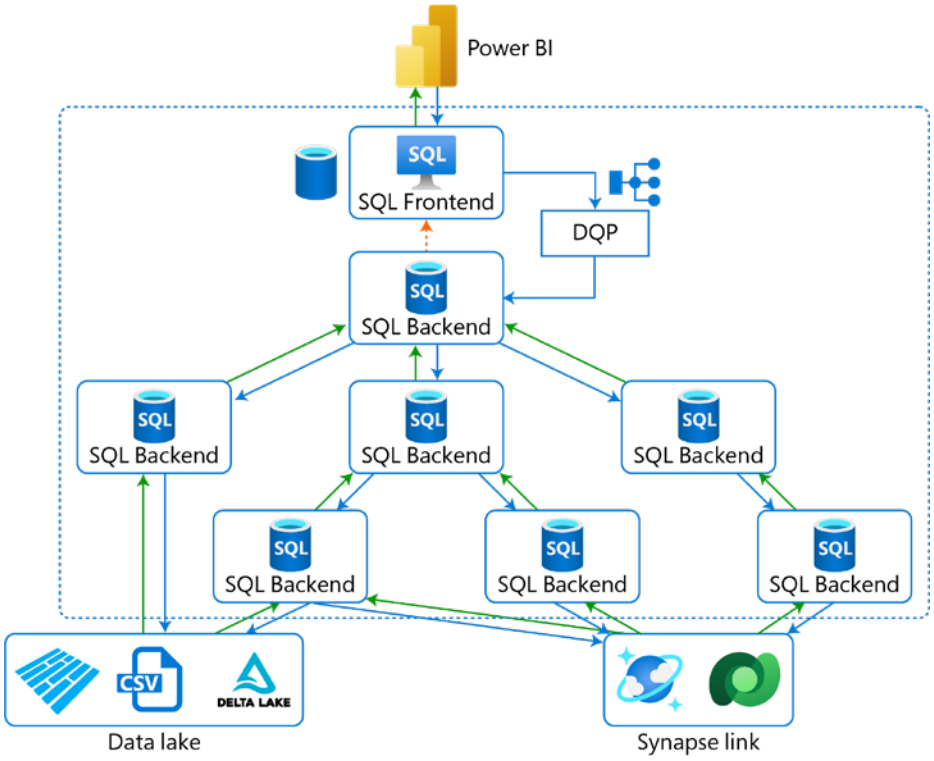

Figure 4: Architecture of Serverless SQL

There are many advantages of using Serverless SQL. You can see a few of them here:

- Easy to discover and explore data in various formats (Delta, Parquet, **Comma-Separated Values** (**CSV**) and **JavaScript Object Notation** (**JSON**)) directly from your data lake.
- Ability to query the analytical storage of your Cosmos DB without impacting the transactional store using Synapse link for Cosmos DB.
- Save money by using the compute only when required.
- No need to worry about infrastructure and managing clusters.
- Easily explore and transform data in a simple, scalable and performant way using T-SQL, and save the results back in a data lake to be visualised further through Power BI reports.
- Build logical data warehouses by providing a relational abstraction on raw data without moving it anywhere. This saves the overhead of additional data ingestion steps and the cost of using Azure resources or any other tool for data movement. However, more importantly, it saves a lot of time by avoiding data movement and trying to keep it updated.

Serverless SQL pool works on the pay-per-query model and, within the **`Manage`** hub of Synapse Studio, you can click on the **`Cost Control`** hyperlink for **`Built-in`** to manage the cost of Serverless SQL.

Azure Synapse Serverless SQL can be the best fit for the **hybrid transactional and analytical processing** (**HTAP**) kind of workload where you can perform analytical operations on data without affecting highly transactional data.

The OPENROWSET function is used in Serverless SQL to query an external data source. This function can be used for reading different types of files including Delta, CSV, JSON and Parquet. Following is one of the examples of OPENROWSET. You can refer to [How to use OPENROWSET using serverless SQL pool in Azure Synapse Analytics] to learn more about using the OPENROWSET function with different file types:

```
select top 10 *
from OPENROWSET(
    bulk 'https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/ecdc_cases/
latest/ecdc_cases.parquet',
    format = 'parquet') as rows
```

Serverless SQL uses OPENROWSET syntax to analyse the data in the analytical store of a Cosmos DB container, as you can see in the following code block:

```
OPENROWSET(
       'CosmosDB',
       '<Azure Cosmos DB connection string>',
       <Container name>
    )  [ < with clause > ]
```

In the next section, we are going to learn how to use Azure Synapse Link to perform streaming analytics on operational data.

5

Operational analytics

Azure Synapse Link allows you to perform analytics on operational data. Azure Synapse Link is a new feature added to create a link between Azure Cosmos DB and Azure Synapse. It enables you to run near-real-time analytics on data residing in the analytical store of your Cosmos DB account. The analytical store and transactional store are kept in sync in a Cosmos DB account. The transactional store in Cosmos DB is optimised for transactional reads and writes, whereas the analytical store is optimised for analytical queries. Synapse Link creates an integration between Cosmos DB and Synapse Analytics.

The analytical store of Cosmos DB can be used to derive analytics on highly transactional data. The following architecture is an example of performing real-time analytics using Azure Synapse Link:

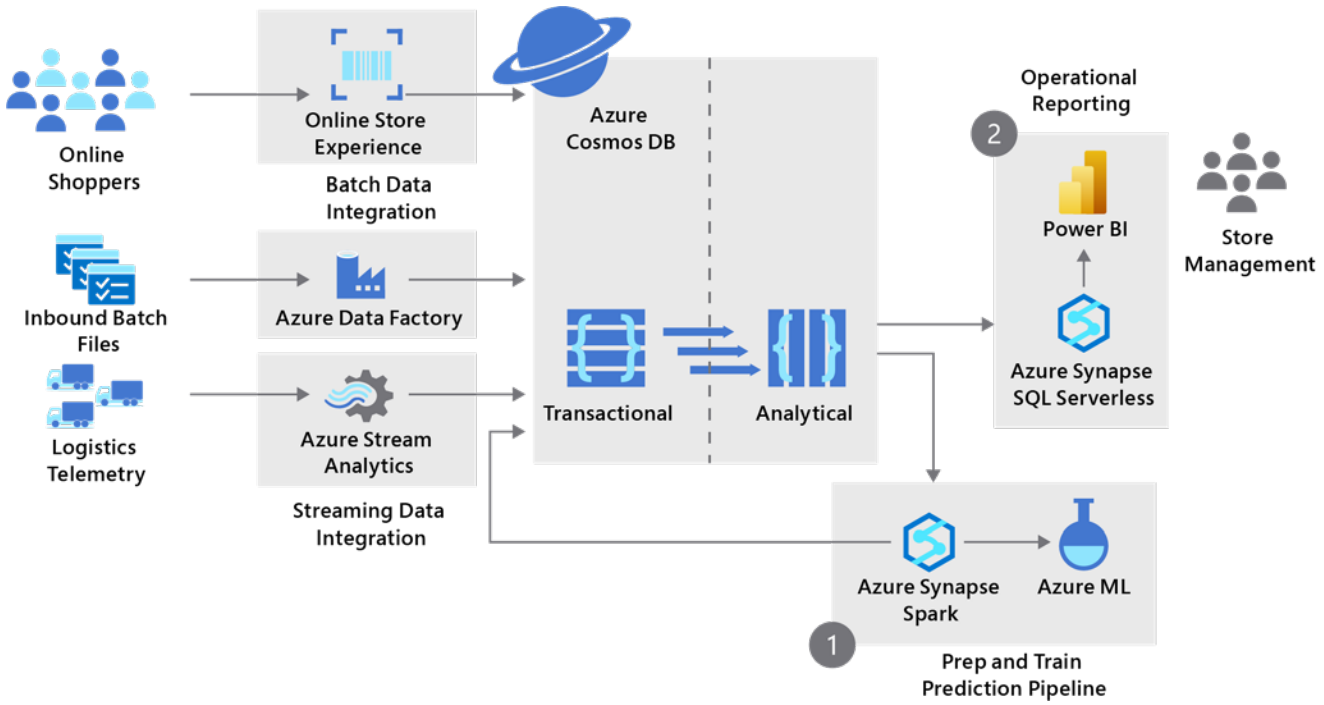

Figure 5: Architecture for real-time analytics on Azure by using Azure Synapse Link

There are various use cases for real-time analytics, including the following:

- **Anomaly detection**: This technique is used to identify unusual behaviour or patterns that raise suspicion because of a significant difference from the rest of the data.
- **Supply chain analytics**: This process is used to increase operational effectiveness by using data and quantitative methods for decision making.
- **Real-time personalisation**: This technique is used to gather information about the user visiting your website and engage that user by providing tailored content on the website based on their company, location, digital behaviour and so on.

In the following section, we will look at data science and predictive analytics.

6

Data science and predictive analytics

Machine Learning (**ML**) has become an integral part of the data ecosystem now and Azure enables you to build powerful, cloud-based ML applications by using the Azure Machine Learning service. Azure ML provides you with options to create supervised or unsupervised ML models and its integration with Azure Synapse has opened a wide ocean for data scientists.

The process of automating the iterative tasks of ML model development is called AutoML. We can build highly scalable, efficient and productive ML models using AutoML. We can use AutoML for classification, regression or forecasting tasks as per our business needs. You can enrich your data with Azure Cognitive Services pre-trained models. Text Analytics (sentiment analysis) and Anomaly Detector are two models available through the Synapse workspace. You can follow the Tutorial to train your first ML model.

Models that have been trained either in Azure Synapse or outside Azure Synapse can easily be used for batch scoring. Currently, in Synapse, there are two ways in which you can run batch scoring:

- You can use the `TSQL PREDICT` function in Synapse SQL pools to run your predictions right where your data lives. This powerful and scalable function allows you to enrich your data without moving any data out of your data warehouse. A new guided ML model experience (Tutorial: Machine learning model scoring wizard (preview) for dedicated SQL pools) in Synapse Studio was introduced where you can deploy an ONNX model from the Azure Machine Learning model registry in Synapse SQL pools for batch scoring using `PREDICT`.
- Another option for batch scoring ML models in Azure Synapse is to leverage Apache Spark pools for Azure Synapse. Depending on the libraries used to train the models, you can use a code experience to run your batch scoring.

To explore all other ML capabilities, go through Machine Learning capabilities in Azure Synapse Analytics.

7

Modernise data warehouse workloads with Synapse and Power BI

Azure Synapse enables you to create your data warehouse using the SQL pools in the cloud. SQL pools can be used to store relational data for running analytical queries against the data at scale.

SQL pools use a scale-out, node-based architecture with one **control node** and multiple **compute nodes** for distributed computational processing. Control nodes are a single point of contact for end users to interact with all compute nodes. The control node runs the MPP engine, which passes an operation to multiple compute nodes to do their work in parallel. MPP databases are optimised for analytical workloads, such as aggregating and processing large datasets. In this type of architecture, each compute node (also known as a processing unit) works independently, with its own operating system and dedicated memory.

Figure 6 represents how all the components are tied together in Azure Synapse SQL pool:

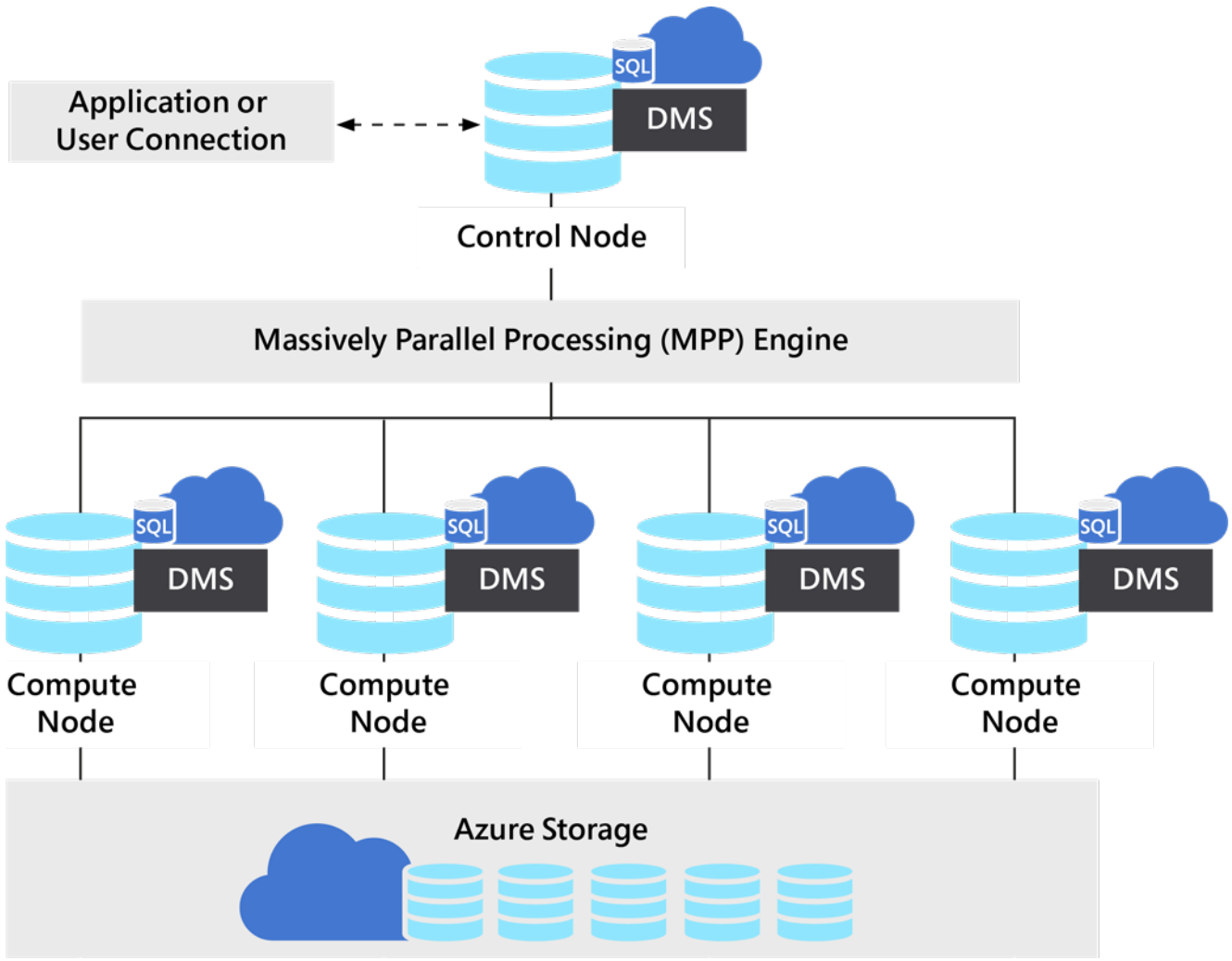

Figure 6: Architecture for SQL pool in Azure Synapse

You can use the following query to get a count of the control and compute nodes available for your Synapse SQL pool:

```
SELECT * FROM sys.dm_pdw_nodes
GO
SELECT type,COUNT(1)
FROM sys.dm_pdw_nodes
GROUP BY type
```

Figure 7 displays the count of **control nodes** and **compute nodes** available in a sample Synapse SQL pool:

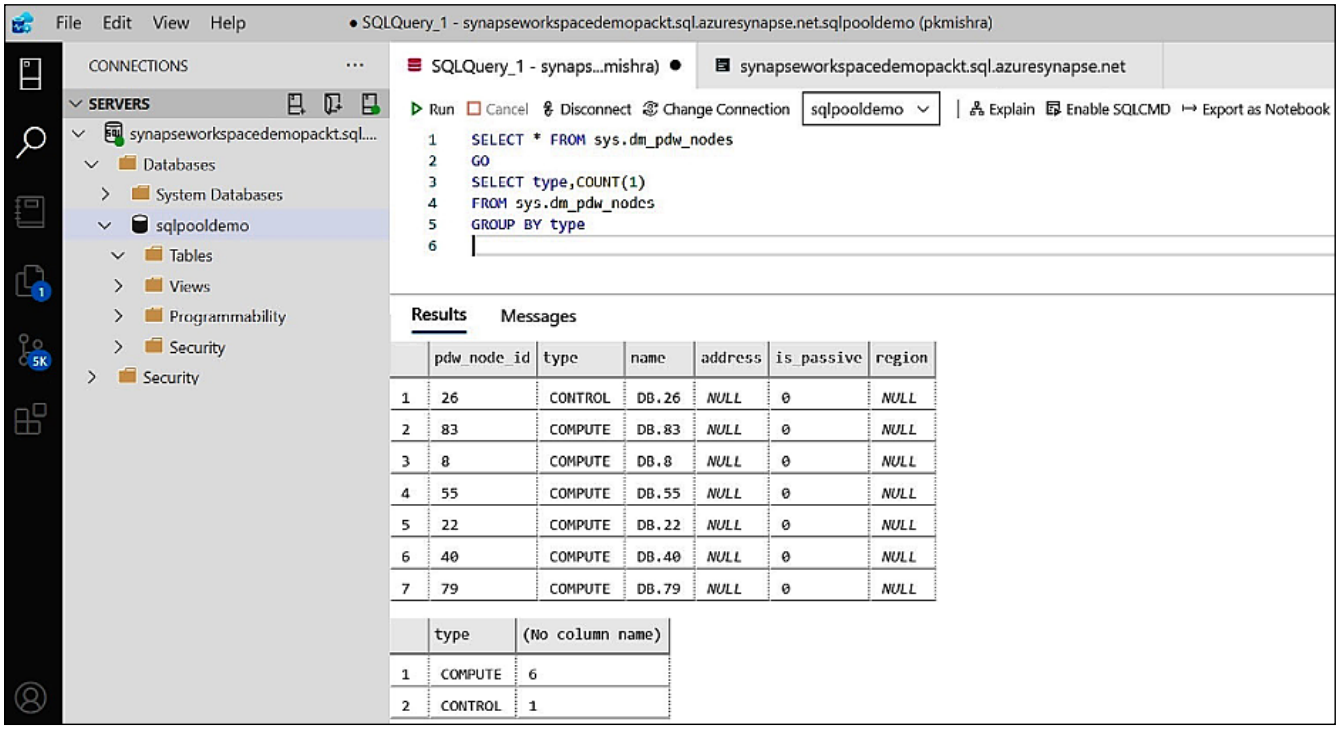

Figure 7: A screenshot of Azure Data Studio showing the query results

When you purchase **Data Warehouse Units** (**DWUs**) for a SQL pool, you basically purchase several analytical resources bundled together, such as the **Central Processing Unit** (**CPU**), memory and **Input/Output** (**I/O**). You can change the DWUs even after creating a Synapse account. However, distributions will remap to compute nodes after you change DWUs for your Synapse account.

You can use the following query to view the current DWU setting:

```
SELECT  db.name [Database]
,         ds.edition [Edition]
,         ds.service_objective [Service Objective]
FROM    sys.database_service_objectives    AS ds
JOIN    sys.databases  AS db ON ds.database_id = db.database_id
```

You can also migrate your on-premises SQL data warehouse to Synapse SQL using Synapse Pathway. You can find more details about Synapse Pathway in Azure Synapse Pathway overview.

Power BI enables you to create models directly from the SQL pool without any need to create a tabular/multidimensional model. You can also create reports directly within Synapse Studio after connecting your Power BI workspace with a Synapse workspace. Reports can be published and shared with other team members to create a collaborative environment.

We will perform the following steps to connect our Synapse workspace to the Power BI workspace:

1. Go to your Synapse workspace and click on the **`Open Synapse Studio`** link, as highlighted in *Figure 8*:

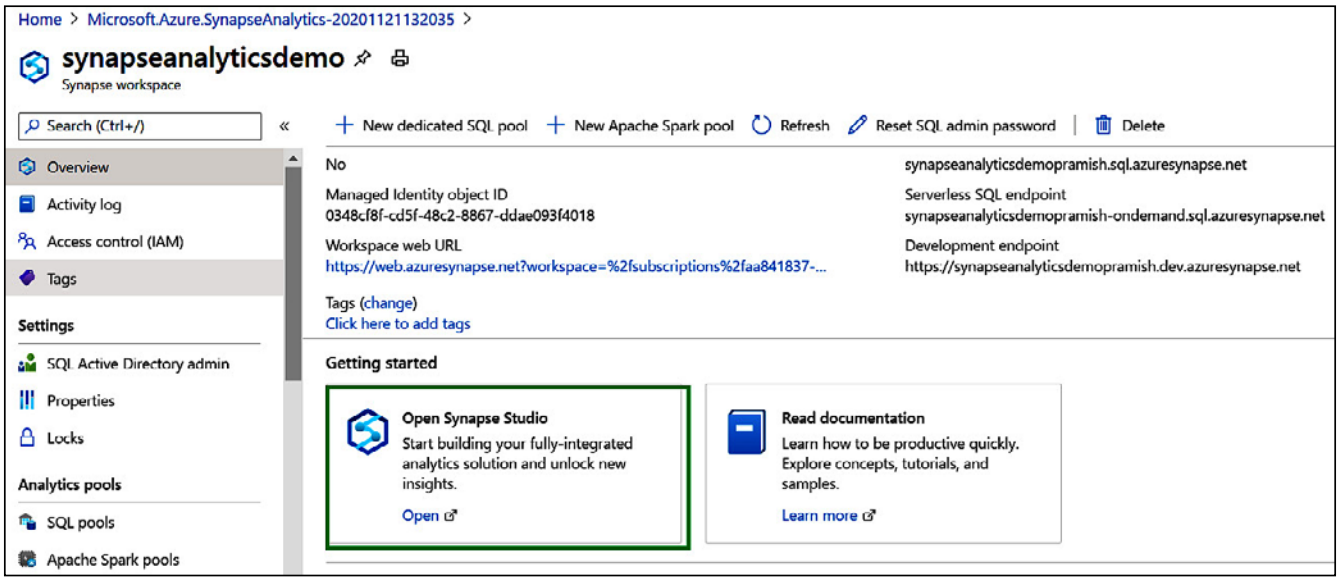

Figure 8: A screenshot of the Azure Synapse workspace highlighting the link to Open Synapse Studio

2. In Synapse Studio, click on the **`Visualize`** tab on the **`Home`** page, as shown in *Figure 9*, or go to the **`Manage`** tab in Synapse Studio and then go to the **`Linked Services`** section to add a new linked service:

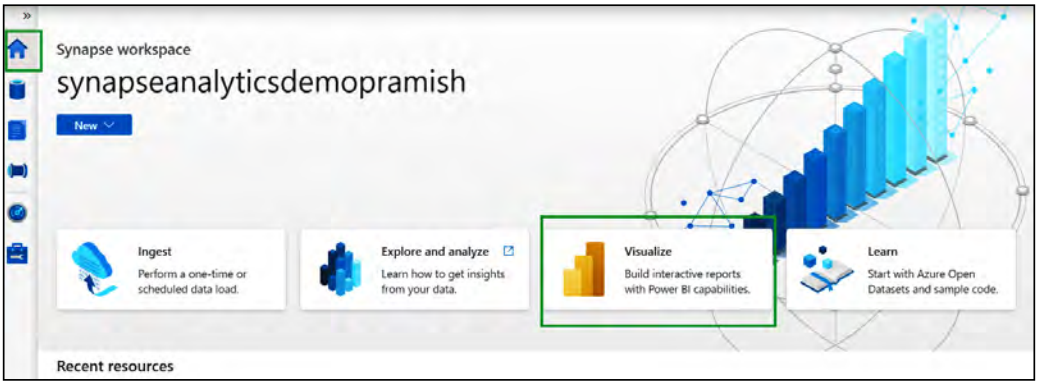

Figure 9: A screenshot of Synapse Studio highlighting the Visualise link

3. Provide an appropriate name and description for the Power BI workspace. This name can be different from the actual Power BI workspace name.
4. Select **`Tenant`** and **`Workspace`** name from the drop-down lists and then click on **`Create`**.

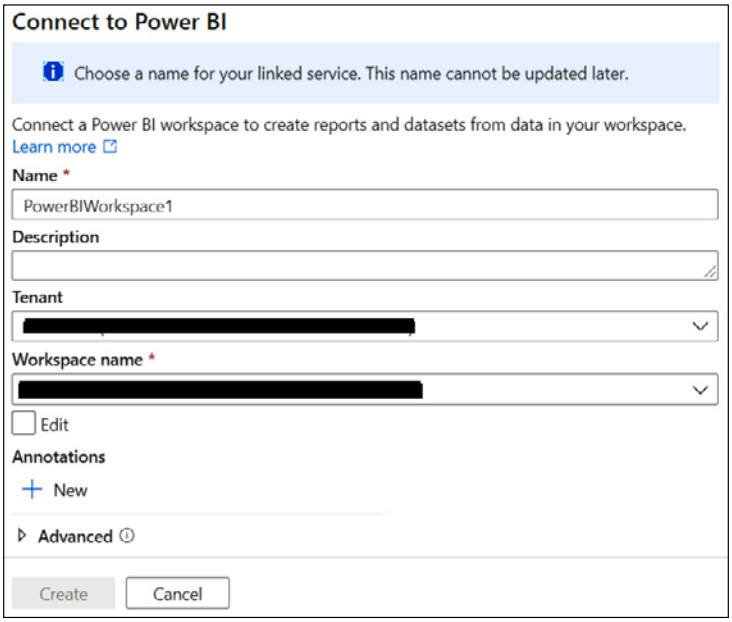

Figure 10: Connecting a Power BI workspace to Azure Synapse

5. Click on the **Develop** tab to verify whether you can see your Power BI workspace under the **Power BI** section. You should be able to see **Power BI datasets** and **Power BI reports** associated with your Power BI workspace.

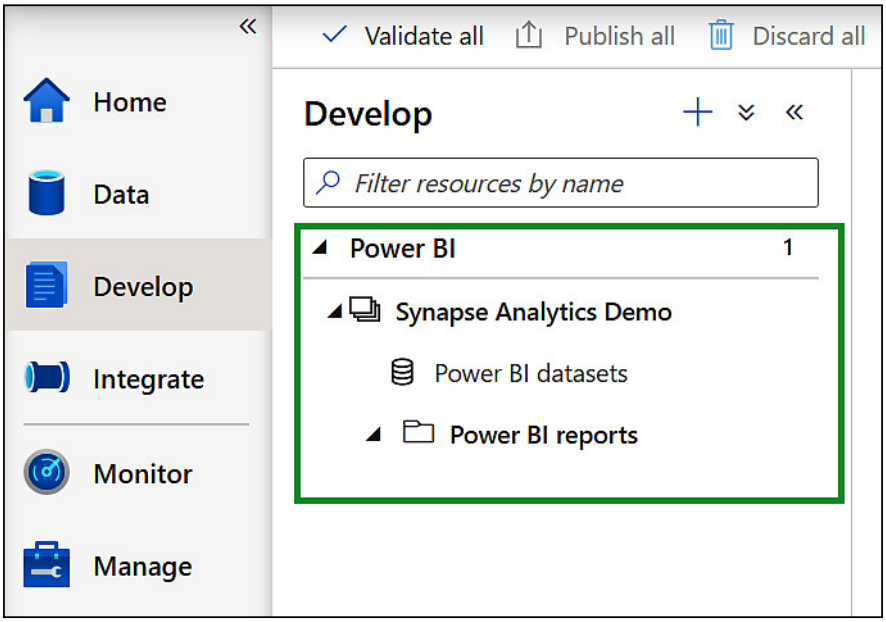

Figure 11: A screenshot of a Power BI workspace within Synapse Studio

Now that you have connected your Synapse workspace to the Power BI workspace, you can create some valuable reports under the **Develop** hub of Synapse Studio to visualise your data.

8

Data governance solution with Azure Purview

Azure Purview is a unified data governance solution for managing on-premises, multi-cloud and Software-as-a-Service data. Azure Purview can help with data management and governance during your data warehouse modernisation journey. The foundation of Azure Purview is the Purview platform. Built on top of the data map are a set of purpose-built data governance applications, including the Data Catalogue (with integrated business glossary) and Data Estate Insights.

Purview Data Map is a cloud-native PaaS service that captures metadata about enterprise data present in analytics and operation systems on-premises and in the cloud. Purview Data Map is automatically kept up to date with a built-in automated scanning and classification system. Purview Data Map powers the Purview Data Catalogue and Purview data insights as unified experiences within the Purview Studio.

Go through Map your data estate with Azure Purview to learn more about Purview Data Map.

With the **Purview Data Catalogue**, business and technical users alike can quickly and easily find relevant data using a search experience with filters based on various lenses, including glossary terms, classifications, sensitivity labels and more. Data consumers and producers can also visually trace the lineage of data assets starting from on-premises operational systems to consumption in an analytics system like Power BI.

To learn more about Purview Data Catalogue:

- Search and browse: Enable effortless discovery of data by business and technical data consumers with Azure Purview.
- Data lineage: Track the lineage of your organisation's data with Azure Purview.
- Glossary: Break free of operational silos with a consistent business glossary.

With **Purview data estate insights**, data officers and security officers can get a bird's-eye view of what data is actively scanned, where sensitive data is and how data moves across systems.

Go through Get a bird's-eye view of your data estate with Azure Purview Data Insights to learn more about Purview Data Map.

You can connect to your Purview account under the **Manage** hub of Synapse Studio:

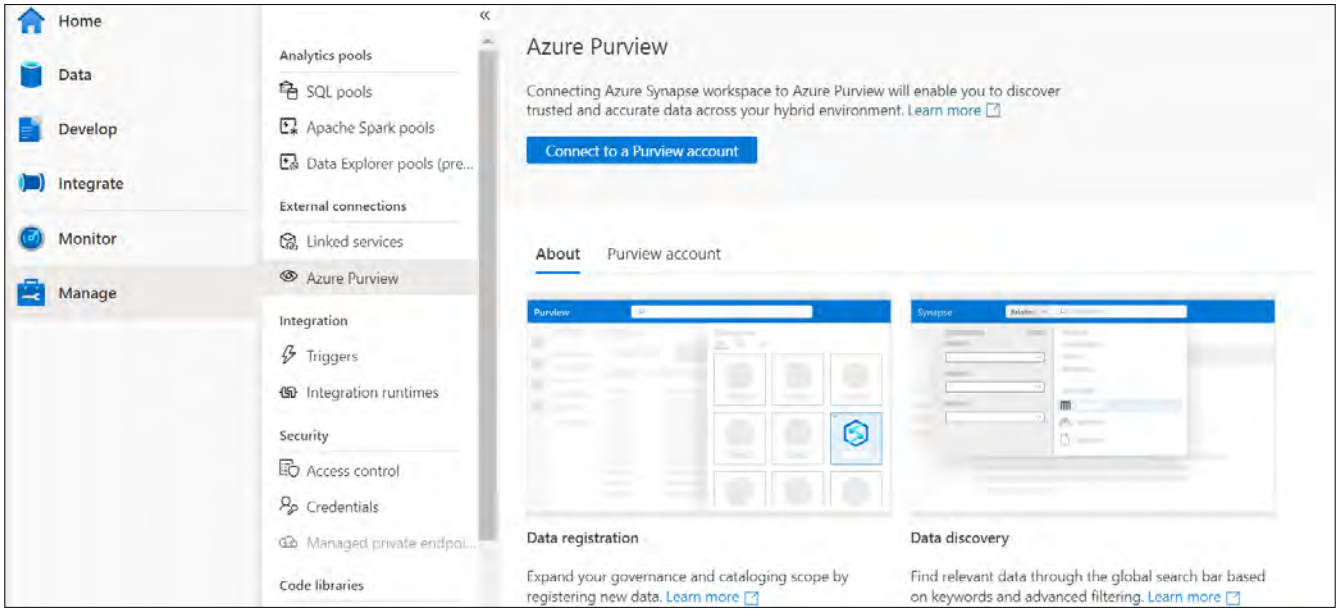

Figure 12: A snapshot of the Manage hub to manage an Azure Purview account within Azure Synapse

Along with data governance, data security is also critical to your workload. We are going to learn about all the security features provided by Azure Synapse in the following section.

9

Advanced security and privacy features

Figure 13 represents the different layers of enterprise-grade security in Synapse. Understanding all these security layers in detail will help us learn the importance of security measures and how we can implement them in our Synapse environment.

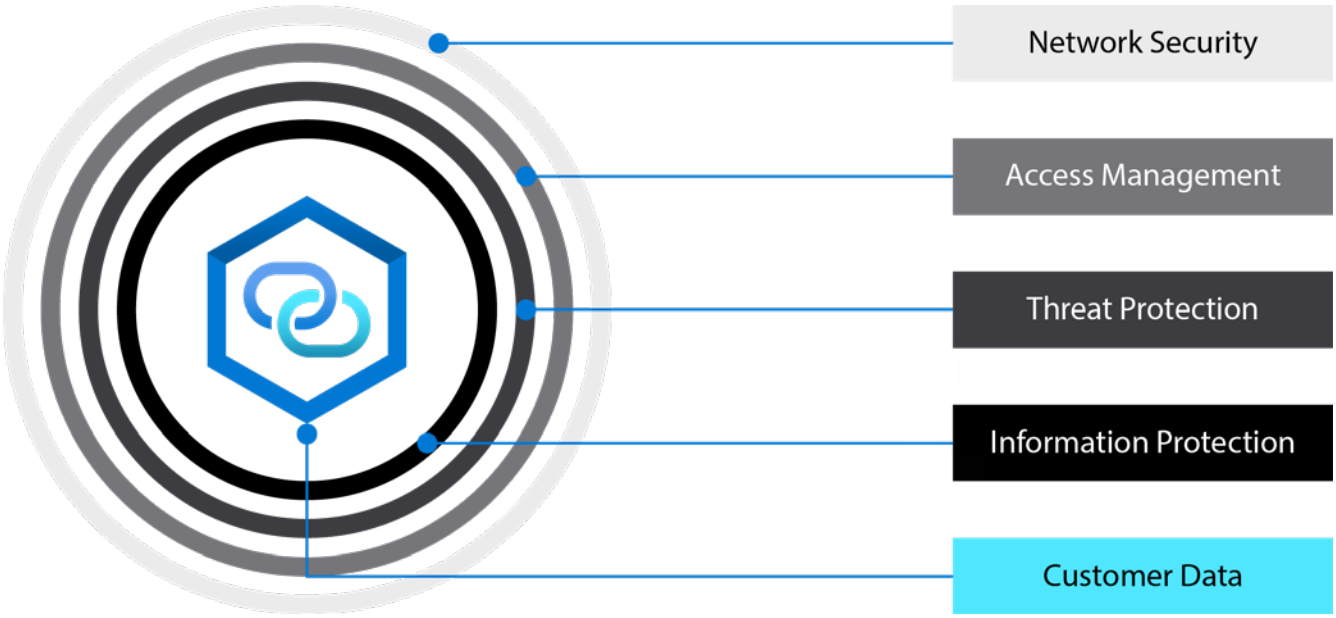

Figure 13: Security layers of Azure Synapse

Network security

Azure Synapse provides you with the option to enable a managed workspace virtual network while creating your Synapse workspace. It ensures that your workspace is isolated from another workspace. If you have enabled a managed workspace virtual network in your Synapse workspace, then data integration and Spark resources are also deployed in the same virtual network; however, SQL pools (dedicated or serverless) reside outside this managed virtual network.

A **private endpoint** is used to provide secure connectivity between your storage and the clients on the virtual network using a private IP address from your virtual network. It is a network interface that enables you to connect to a service securely powered by Azure Private Link. You can go through What is Azure Private Endpoint? to understand Private Link in detail.

IP firewall rules enable you to access SQL pools from the IP addresses that are whitelisted in the IP firewall rules.

Access management

Azure Synapse provides a comprehensive and fine-grained access control system that integrates:

- **Azure roles** for resource management and access to data in storage
- **Synapse roles** for managing live access to code and execution
- **SQL roles** for data plane access to data in SQL pools
- **Git permissions** for source code control, including continuous integration and deployment support

Azure Synapse roles provide sets of permissions that can be applied to varying extents. This granularity makes it easy to grant appropriate access to administrators, developers, security personnel and operators to compute resources and data.

Access control can be simplified by using security groups that are aligned with people's job roles. You only need to add and remove users from appropriate security groups to manage access.

Threat protection

It is important to protect our data from any anomalous activities that could be potentially harmful attempts to exploit our databases. Synapse provides you with two ways to protect your data against any threat. The first one is **SQL auditing**, which captures the activities related to all the changes to security, access to tables and many more activities besides, to protect your data. The second is **Azure Defender**, which checks the vulnerability of your SQL pools and provides advanced data security for your data.

Azure SQL auditing captures all the events in a Synapse SQL pool and writes them to an audit log in your Azure Storage account. These audit logs can be used to analyse anomalous activities or unexpected behaviour in the SQL pool.

This feature will be disabled by default, but you can enable it on the **Azure SQL Auditing** tab of your Azure Synapse workspace.

Information protection

Sometimes, just storing data securely is not sufficient. We need to protect data even when it is in motion and in use. Azure provides different security features to protect your data at any given time so that you can meet all the data-related compliances. These are as follows:

- **Encryption-in-flight (Transport Layer Security – TLS)**: The Synapse SQL pool secures your data by encrypting data in motion with TLS.
- **Encryption-at-rest (Transparent Data Encryption – TDE)**: TDE encrypts your databases, backups and logs at rest. This setting is specific to one particular SQL pool. If you create another SQL pool in your Synapse workspace, then you need to enable this setting separately for that pool.

Figure 14 shows how to enable data encryption for your Synapse SQL pools:

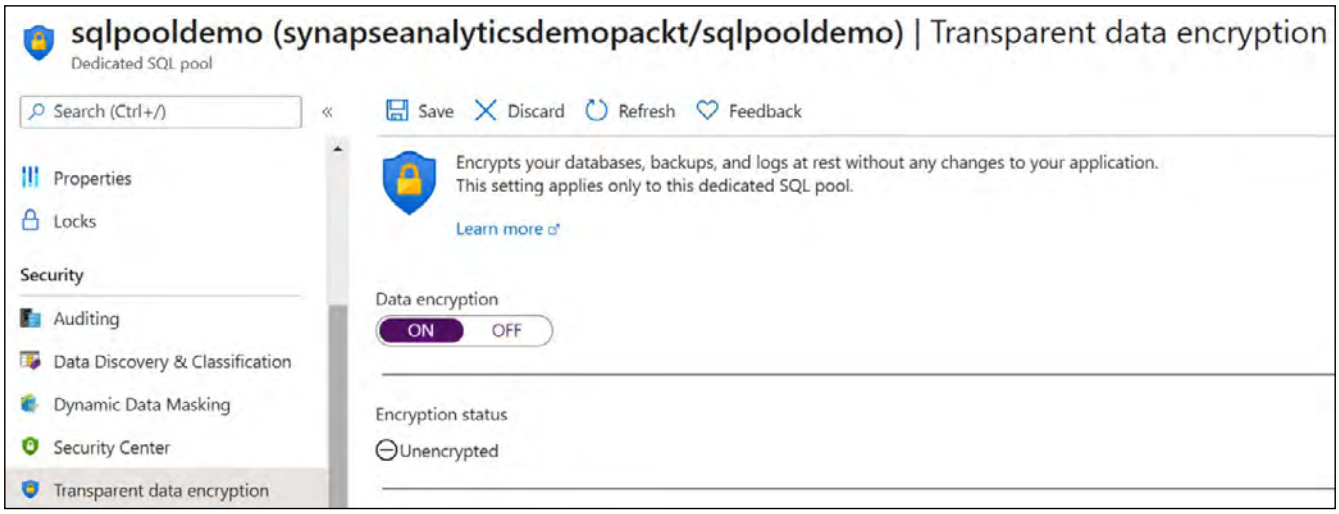

Figure 14: Enabling TDE for the dedicated SQL pool in Azure Synapse

- **Encryption-in-use (always encrypted)**: The always encrypted feature is designed to protect sensitive data; it only makes the data available to client-side applications, and the data will not be visible to administrators either.

In this section, we got to learn how to implement network security, access management, threat protection and information protection. However, you can follow [Azure Synapse Analytics security white paper: Introduction] to learn more about securing your Synapse workspace.

10

Save on costs with Azure Synapse

There is always an advantage in terms of cost when you move from on-premises to PaaS. You can leverage some of the cost-saving options listed below for your Synapse workspace:

- Use Serverless SQL to use the compute as per requirements instead of having persistent compute.
- Configure the **`Auto-pause`** setting of a Spark pool to pause a cluster automatically if not in use.
- Use the built-in web-based managing and monitoring tool provided within Synapse Studio.
- Save costs for Azure Synapse Analytics charges with reserved capacity.

Conclusion

Azure Synapse is a one-stop platform for managing your analytical workload at scale. It provides you with a suite of data-related services under one unified experience that makes it very easy and convenient to manage your analytical workload without any hiccups. Azure Synapse Pathway paves an easier path for you to move your on-premises data warehouse objects to Azure Synapse with just a couple of clicks. Apart from the SQL workload, you can also run your R or Python script against the data stored in your data lake without moving the data anywhere.

To learn more:

- Read the full version: **Limitless Analytics with Azure Synapse**
- Get started on Azure Synapse Analytics with **an Azure account**
- Visit the **Azure Synapse Analytics documentation web page** for tutorials
- Request a call from an **Azure Synapse Analytics sales specialist** when you're ready