Building Data Pipelines in Azure with Azure Synapse Analytics

CookBook

Table of Content:

Introduction

In this training, we shall look into a pipeline that serves historical data with the highest value as the data point is generated, corresponding to which an actionable item needs to be initiated. Before the development of Azure Synapse Analytics, organizations faced several challenges when processing and analysing big data. Some of the main challenges include:

- **Data silos:** Many organizations had data stored in multiple different systems and databases, making it difficult to access and analyse the data in a meaningful way.
- **Complex data processing:** Big data processing tasks, such as data ingestion, transformation, and machine learning, can be complex and time-consuming, requiring specialized skills and hardware.
- **Inadequate data warehousing:** Traditional data warehousing solutions were not designed to handle the volume, variety, and velocity of big data, making it difficult for organizations to store and analyse this data effectively.
- **Inefficient data analytics:** Organizations often struggled to extract insights from their data, due to the complexity of the data and the lack of appropriate tools and platforms.
- **High costs:** The cost of hardware and infrastructure to support big data processing and data warehousing can be prohibitively expensive for many organizations.

By developing Azure Synapse Analytics, Microsoft aimed to address these challenges by providing a unified and seamless solution for big data analytics and data warehousing, making it easier and more affordable for organizations to process, store, and analyze their big data. With Azure Synapse Analytics, organizations can benefit from a cloud-based platform that can handle their big data needs and provide insights more quickly and easily. Before moving on to the actual pipeline, let's see the difference between traditional and Big data.

Traditional data:
Traditional data refers to the data that has been collected, processed, and analyzed using traditional methods and technologies, such as relational databases and data warehousing. The main characteristics of traditional data are its limited volume, structured format, and centralized storage.

A traditional data source is a data source that is based on structured data, typically stored in databases or spreadsheets. Examples of traditional data sources include:

1. **Relational databases:** These are structured databases that store data in tables with defined relationships between the tables.
2. **Flat files:** These are simple text files that contain data in a tabular format.
3. Spreadsheets: These are electronic documents that contain data organized in rows and columns, and can be used for a wide range of data analysis and visualization tasks.
4. **Legacy systems:** These are older computer systems that have been in use for a long time, and often contain data that is not stored in a modern format.
5. **Operational systems:** These are systems that are used to support day-to-day business operations, such as enterprise resource planning (ERP) or customer relationship management (CRM) systems.

These traditional data sources are still widely used today, and provide a stable and well-understood foundation for data storage and analysis. However, as the volume, variety, and velocity of data have increased, many organizations are looking for more advanced solutions to handle big data.

Big data:
Big data, on the other hand, refers to data that exceeds the processing capacity of traditional database systems due to its massive volume, high velocity, and complex structure. Big data is typically generated from various sources, such as social media, sensors, and mobile devices, and it often requires advanced technologies, such as distributed systems and machine learning algorithms, to process and analyse it effectively. The main characteristics of big data are its large volume, high velocity, and diverse structure.

A big data source is a data source that generates a large volume of data, often in a variety of formats and at high speed. Examples of big data sources include:

1. **Social media:** Sites such as Twitter, Facebook, and Instagram generate vast amounts of data in real-time, including text, images, and videos.
2. **IoT devices:** The Internet of Things (IoT) refers to a network of connected devices that can generate large amounts of data, such as temperature readings, sensor data, and location data.
3. **Weblogs:** Web logs are records of all activity on a website, including page views, clicks, and other interactions.
4. **E-commerce transactions:** Online stores and marketplaces generate data on customer purchases, shipping, and other transactions.
5. **Video and audio:** Streaming services, such as YouTube and Netflix, generate large amounts of video and audio data.

These big data sources are often characterized by the 3 Vs of big data: volume, velocity, and variety. To effectively analyse this data, specialized tools and techniques are needed, such as distributed computing systems, data lakes, and machine learning algorithms.

Here is a table summarizing the differences between traditional data and big data:

Characteristic	Traditional Data	Big Data
Volume	Limited	Large
Velocity	Low	High
Structure	Structured	Semi-structured/ Unstructured
Processing	Traditional methods and technologies	Advanced technologies
Analysis	Predetermined and rule-based	Predictive and exploratory
Storage	Centralized	Decentralized

This table highlights the main differences between traditional data and big data, such as the volume of data, the speed at which it is generated, the format of the data, the methods and technologies used for processing and analysing the data, and the storage methods used.

Did you know?

There is no single "hidden fact" about big data, as it is a complex and multifaceted field with many different aspects to consider. However, some common misconceptions or lesser-known aspects of big data include:

1. *Big data is not just about quantity, but also about the quality and variety of data.*
2. *Big data can contain a large amount of unstructured data, such as text, images, and videos, which can be difficult to analyse.*
3. *Big data can also generate ethical and privacy concerns, as large amounts of personal information can be collected, stored, and analysed.*
4. *The ability to effectively utilize big data requires not only large amounts of storage and computing power, but also specialized skills, such as data science and machine learning expertise.*
5. *Big data can lead to improved decision making, but it is not a panacea and must be used in conjunction with human expertise and judgment to ensure accurate and ethical results.*

Azure Synapse Analytics

In this project, you will learn how to build a data pipeline using Azure Synapse Analytics. Before we dive deep into the actual pipeline, let's understand Azure Synapse Analytics in detail.

Azure Synapse Analytics is a big data and data warehousing solution offered by Microsoft as part of the Azure cloud platform. It integrates big data and data warehousing into a single service, making it easier for organizations to analyze and manage large amounts of data.

The key features of Azure Synapse Analytics include:

1. **Data lake storage:** Azure Synapse Analytics includes a data lake that can store a large volume of structured and unstructured data, including data from IoT devices, log files, and social media.
2. **Spark:** This component provides a distributed computing framework for processing big data in real-time, and is integrated into Azure Synapse Analytics to enable organizations to perform complex data processing tasks.
3. **Analytics services:** The service includes built-in analytics tools, such as Spark and Azure Machine Learning, to enable organizations to perform complex data processing and analysis tasks.
4. **Integration with Azure Data Warehouse:** Azure Synapse Analytics integrates with Azure Data Warehouse, a cloud-based data warehousing solution, to provide organizations with a single platform for both big data and data warehousing.
5. **Power BI integration:** Azure Synapse Analytics can be integrated with Power BI, a data visualization and business intelligence tool, to provide organizations with powerful insights into their data.
6. **Security and privacy:** Azure Synapse Analytics include robust security and privacy features, including encryption at rest and in transit, to ensure the confidentiality and integrity of sensitive data.
7. **Data factory:** This component provides a cloud-based data integration and management solution that enables organizations to move and transform data from a variety of sources.
8. **Data catalog:** This component provides a centralized metadata repository that makes it easier for organizations to discover and understand the data stored in their data lake.
9. **Workspace:** This component provides a single interface for accessing all of the components of Azure Synapse Analytics, making it easier for organizations to manage and analyze their data.

Overall, Azure Synapse Analytics is designed to help organizations manage and analyze big data in a seamless and scalable manner, making it easier to gain valuable insights from their data and make informed decisions.

Azure Synapse Pools

Pools in Azure Synapse Analytics are virtual computing environments that provide organizations with the ability to run big data analytics and data warehousing workloads. Pools enable organizations to manage their data and analytics tasks within a single workspace, without having to manage infrastructure.
There are two main types of pools in Azure Synapse Analytics:

1. **SQL pool:** This is a relational database service that provides organizations with the ability to store and analyze structured data using the familiar SQL language. It provides a high-performance, scalable, and secure data warehousing solution that integrates with Azure Synapse Analytics.
2. **Spark pool:** This is a virtual computing environment that is optimized for running big data analytics and processing workloads using Apache Spark. It provides organizations with the ability to run Spark jobs and perform complex data processing tasks within the same workspace as their structured data.

Each pool provides organizations with different capabilities for managing and analyzing data in Azure Synapse Analytics. The SQL pool provides a high-performance and scalable data warehousing solution, while the Spark pool provides the computing power needed to process big data in real time. By using these pools in combination, organizations can easily manage and analyze both structured and unstructured data, gaining valuable insights and making informed decisions.

Here's a summary of a comparison table between the SQL pool and Spark pool in Azure Synapse Analytics:

Feature	SQL pool	Spark pool
Purpose	Structured data warehousing and analysis	Big data analytics and processing
Language	SQL	Apache Spark
Scalability	Scalable and high-performance	Scalable and high-performance
Data Types	Structured data	Structured and unstructured data
Processing Speed	Fast for structured data queries	Fast for big data processing
Integration	Integrates with other Azure services, such as Power BI and Azure Machine Learning	Integrates with other Azure services, such as Power BI and Azure Machine Learning
Cost	Pay-as-you-go, only pay for the resources you use	Pay-as-you-go, only pay for the resources you use

Data Description

The Tokyo 2021 Olympics dataset refers to the collection of data related to the Summer Olympic Games that took place in Tokyo, Japan in 2021. This data can come from a variety of sources, including sports organizations, government agencies, media outlets, and other sources.
The data typically includes information on various aspects of the Olympic Games, such as the different events, athletes, and countries participating, as well as results and medal counts.

The data in the Tokyo Olympics dataset can be used for various purposes, including:

- Analysing the performance of different countries and athletes
- Identifying trends in Olympic sports
- Comparing the performance of athletes from different countries and regions
- Exploring the relationship between various factors, such as age, gender, and nationality, and athletic performance

This data can be used by researchers, sports organizations, media outlets, and other organizations to better understand the impact of the games, the performance of the athletes, and the overall experience of the games. It can also be used to help plan future Olympic Games and other large sporting events.

In the age of big data, the dataset for the Tokyo 2021 Olympics is likely to be massive, and may include data from various sources such as social media, sensor networks, and other sources. This data can be analysed using various big data tools and techniques, such as machine learning and data visualization, to gain valuable insights and make informed decisions.

The data includes information on more than 11,000 athletes competing in 47 sports for 743 Teams in the Tokyo Olympics in 2021. This dataset includes information on the participating Teams, Athletes, Coaches, and Entries by gender. It includes their names, nationalities, sports they compete in, and names of coaches. The dataset contains 5 files as follows:

1. **Athletes file:**

   The athletes file in the Tokyo Olympics dataset is a data file that contains information about the athletes participating in the 2020 Summer Olympics held in Tokyo, Japan. The information in the athletes file may include details such as the athlete's name, nationality, and discipline.

PersonName	Country	Discipline
AALERUD Katrine	Norway	Cycling Road
ABAD Nestor	Spain	Artistic Gymnastics
ABAGNALE Giovanni	Italy	Rowing
ABALDE Alberto	Spain	Basketball

2. **Coaches file:**
The coaches file in the Tokyo Olympics dataset is a data file that contains information about the coaches of the athletes participating in the 2020 Summer Olympics held in Tokyo, Japan. The information in the coaches file may include details such as the coach's name, nationality, and discipline.

Name	Country	Discipline
ABDELMAGID Wael	Egypt	Football
ABE Junya	Japan	Volleyball
ABE Katsuhiko	Japan	Basketball
ADAMA Cherif	Côte d'Ivoire	Football

3. **EntriesGender file:**
The EntriesGender file is an important part of the Tokyo Olympics dataset, as it provides information on the gender distribution of athletes participating in the Olympic Games. This information can be used to analyze the representation of men and women in different sports and events, as well as to explore trends and patterns in gender representation over time.

Discipline	Female	Male	Total
3x3 Basketball	32	32	64
Archery	64	64	128
Artistic Gymnastics	98	98	196
Artistic Swimming	105	0	105
Athletics	969	1072	2041
Badminton	86	87	173

4. **Medals file:** (Contains the Medals and Scoreboard of countries that participated in Olympics):

Rank	Team_Country	Gold	Silver	Bronze	Total	Rank by Total
1	United States of America	39	41	33	113	1
2	People's Republic of China	38	32	18	88	2
3	Japan	27	14	17	58	5
4	Great Britain	22	21	22	65	4
5	ROC	20	28	23	71	3

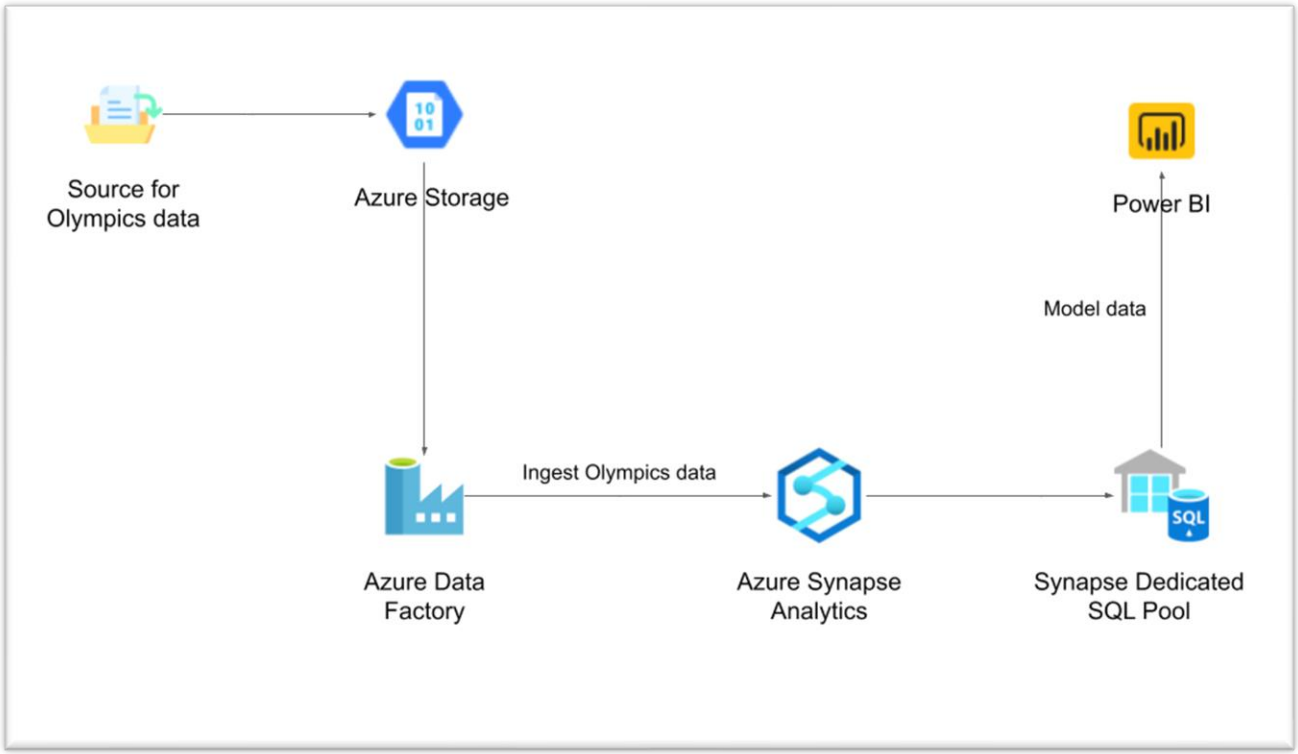
5. **Teams file:** (Details about the Teams, discipline, Name of Country and the event):

TeamName	Discipline	Country	Event
Belgium	3x3 Basketball	Belgium	Men
China	3x3 Basketball	People's Republic of China	Men
China	3x3 Basketball	People's Republic of China	Women
France	3x3 Basketball	France	Women
Italy	3x3 Basketball	Italy	Women
Japan	3x3 Basketball	Japan	Men
Japan	3x3 Basketball	Japan	Women

Use-case depicted in this project

In this project, you will learn how to build a data pipeline using Azure Synapse Analytics, Azure Storage, and Azure Synapse SQL pool to analyze the 2021 Olympics dataset. In this training, we will use the Azure Synapse Dedicated SQL pool. In the next project, we will use the Azure Synapse Spark pool to implement the same pipeline.

The **Architecture** of the pipeline is as follows:



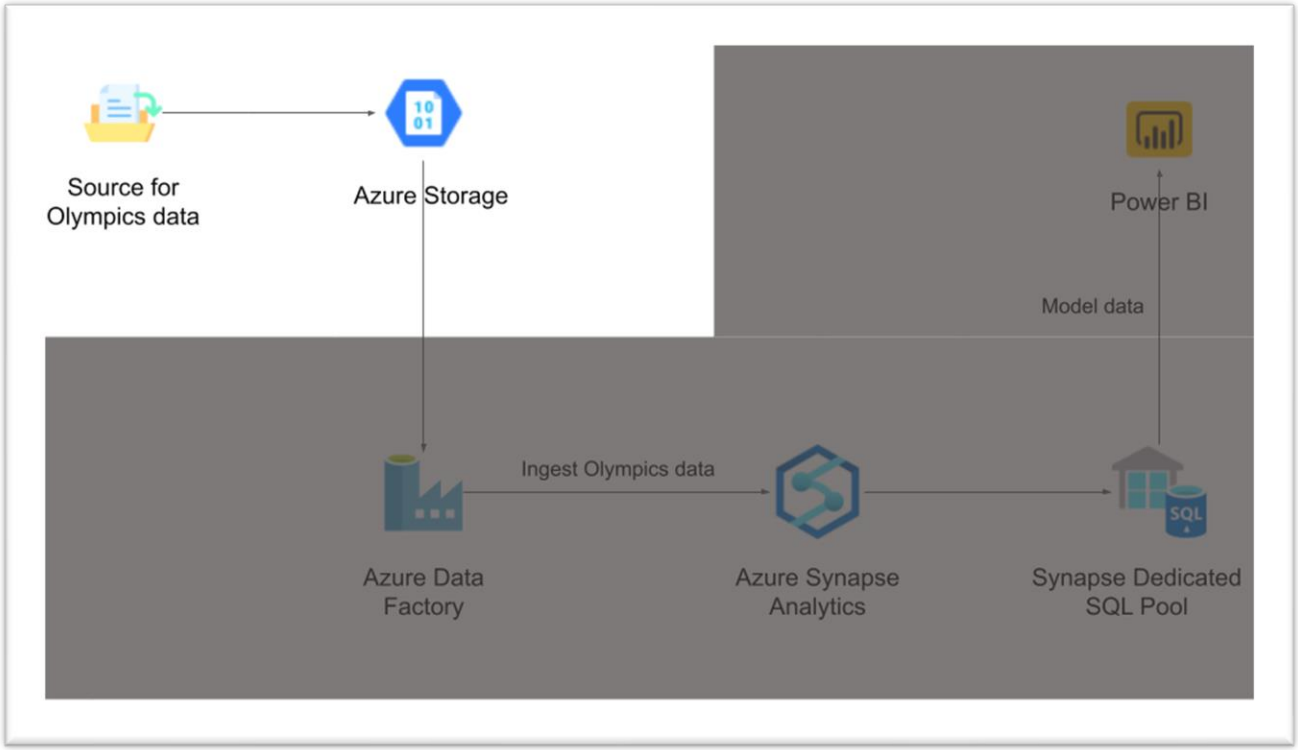
Breakdown of the Use-Case
Setup required:

- Create a Microsoft Azure account
- Download the dataset and the Code files from the project
- This is an OS-independent project, as the whole pipeline is set up on Azure Cloud

Project Workflow:

1. Create an azure storage account and upload data files in a container.
2. Create an azure synapse analytics workspace.
3. Create a SQL pool in azure synapse workspace.
4. Create table structure in SQL pool.
5. Create a data pipeline to ingest data from azure storage into SQL pool tables.
6. Load data from SQL pool tables into Power BI.
7. Prepare dashboard in Power BI.
8. Publish Power BI dashboard in Azure synapse workspace.

Loading data to Azure Blob Storage



Azure Blob storage is a highly scalable, cloud-based object storage service provided by Microsoft Azure. It is used for storing unstructured data, such as images, videos, audio, and text files, as well as binary data.

Blob storage provides a flexible and cost-effective way to store large amounts of data that can be accessed from anywhere in the world via HTTP or HTTPS. Blob storage is also highly scalable, allowing organizations to store and access an unlimited amount of data.

Blobs can be stored in containers, which are logical groupings of blobs within a storage account. A storage account can contain multiple containers, and each container can store an unlimited number of blobs. Blobs are accessed by their URL, which is based on the storage account name and the container name.

Blob storage offers several options for data redundancy, including local redundancy, geo-redundant storage (GRS), and read-access geo-redundant storage (RA-GRS). This provides organizations with the ability to store their data in multiple locations for disaster recovery purposes.

In addition to its scalability and data redundancy options, Blob storage provides several security features, such as encryption at rest and in transit, role-based access control, and network isolation.

Blob storage can be accessed using a variety of methods, including the Azure Portal, Azure CLI, Azure Storage REST API, and Azure Storage client libraries. The Azure Storage REST API provides a way to access and manage Blob storage programmatically, while the Azure Storage client libraries provide a higher-level, more convenient way to access Blob storage.

Overall, Azure Blob storage is a comprehensive and flexible cloud-based storage solution that can meet the needs of organizations of all sizes, from small businesses to large enterprises, for storing and managing large amounts of unstructured data in the cloud.

Here are the steps to create an Azure Blob storage account.

Create an Azure Storage account:

- Go to the Azure Portal and sign in with your Microsoft account.
- Click the "Create a resource" button.
- Search for "Storage account" and click on the result to start the creation process.
- Fill in the required fields for the storage account, such as the name, subscription, resource group, and location.
- Choose "Blob storage" as the account kind.
- Choose the appropriate redundancy options.
- Click the "Review + create" button to review your settings and create the storage account.

Here are the steps to create a container in Azure Blob Storage and upload files in the container:

Create a container:

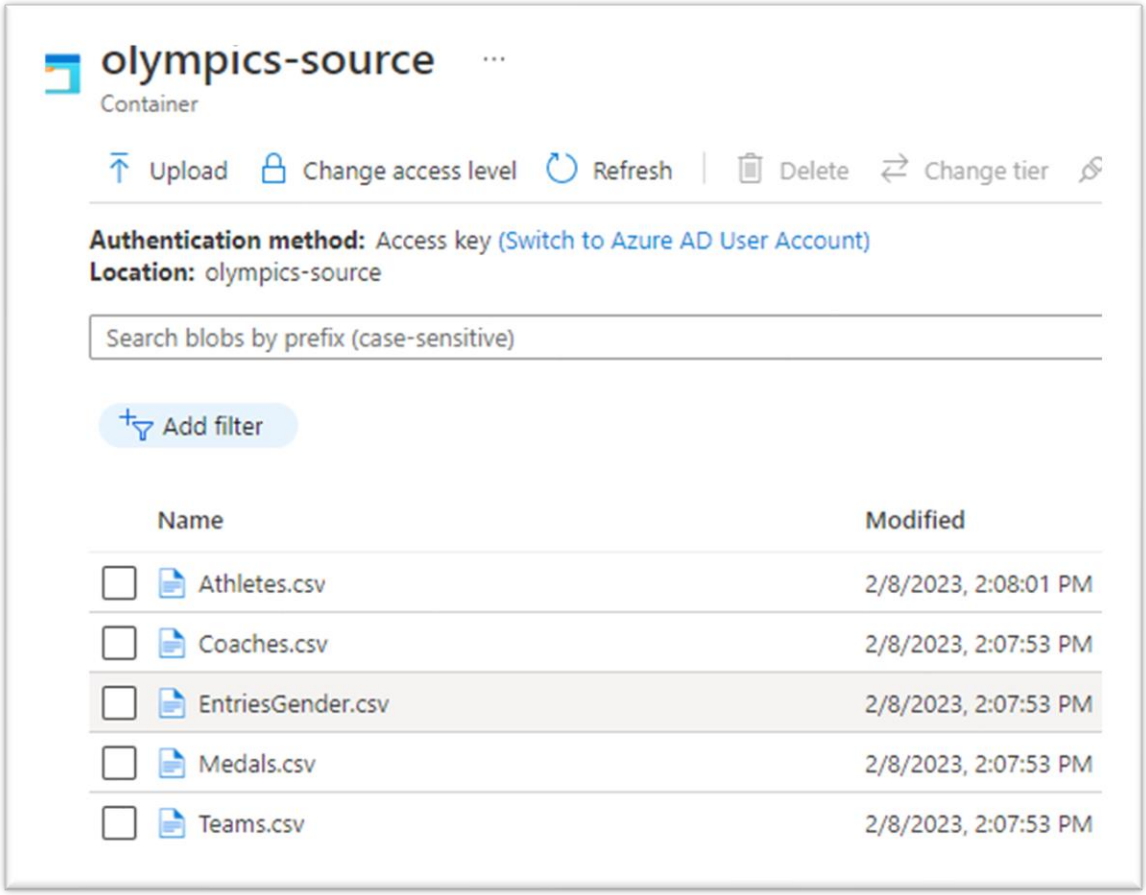
- Go to the Azure Portal and sign in with your Microsoft account.
- Go to your newly created Storage account.
- Click on the "Containers" section.
- Click on the "+ Container" button to create a new container.
- Give your container a name, and choose the appropriate access level.
- Click the "Create" button to create the container.

Upload a CSV file:

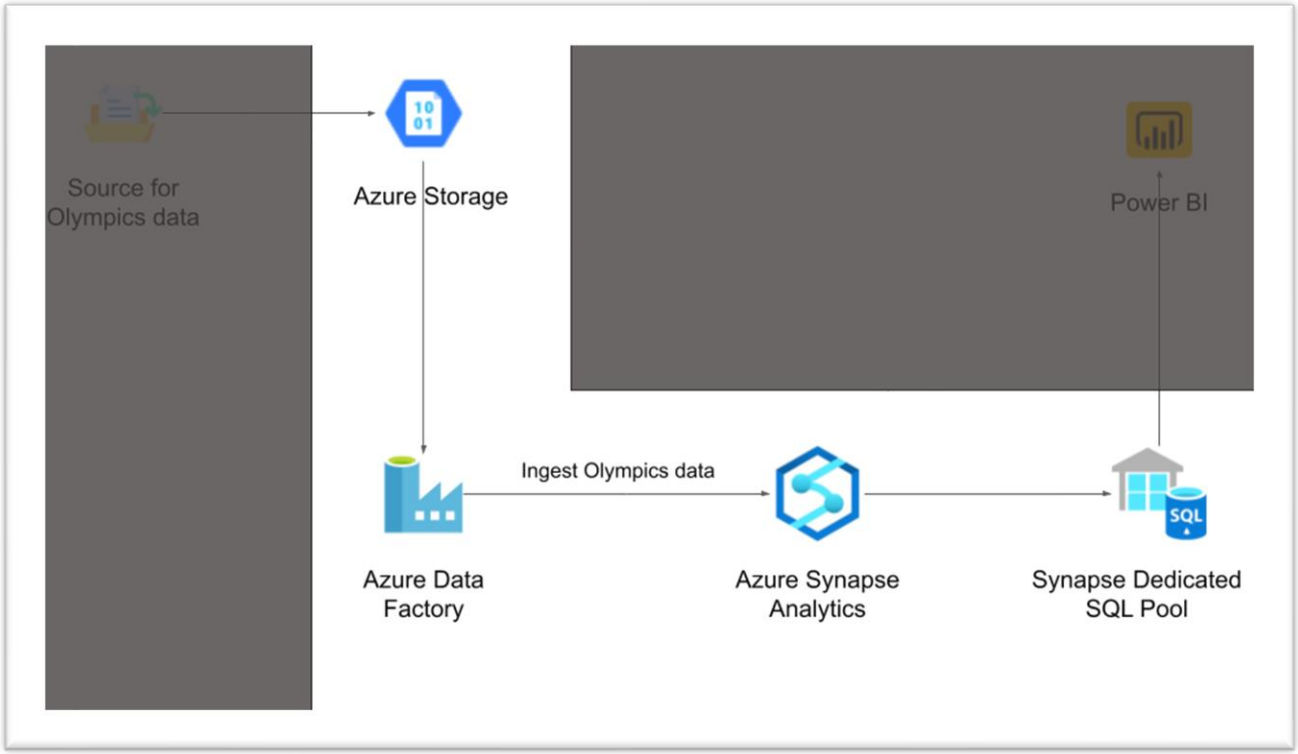
You can use Azure Storage Explorer, the Azure Portal, or the Azure CLI to upload your CSV file.

1. To upload the file using Azure Storage Explorer:
    - Install Azure Storage Explorer from https://azure.microsoft.com/en-us/features/storage-explorer/
    - Connect to your storage account in Azure Storage Explorer.
    - Right-click on the container you created earlier and select "Upload file".
    - Select the CSV file you want to upload and click "Open".
    - Wait for the upload to complete, then check the container to verify the file is there.
2. To upload the file using the Azure Portal:
    - Go to the Azure Portal and sign in with your Microsoft account.
    - Go to your newly created Storage account.
    - Click on the "Containers" section.
    - Click on the container you created earlier.
    - Click the "Upload" button.
    - Select the CSV file you want to upload and click "Open".
    - Wait for the upload to complete, then check the container to verify the file is there.

After uploading all the data files, the container in Azure Storage will look like this:



Creating Pipeline in Azure Synapse to ingest data from Azure Storage to SQL Pool tables



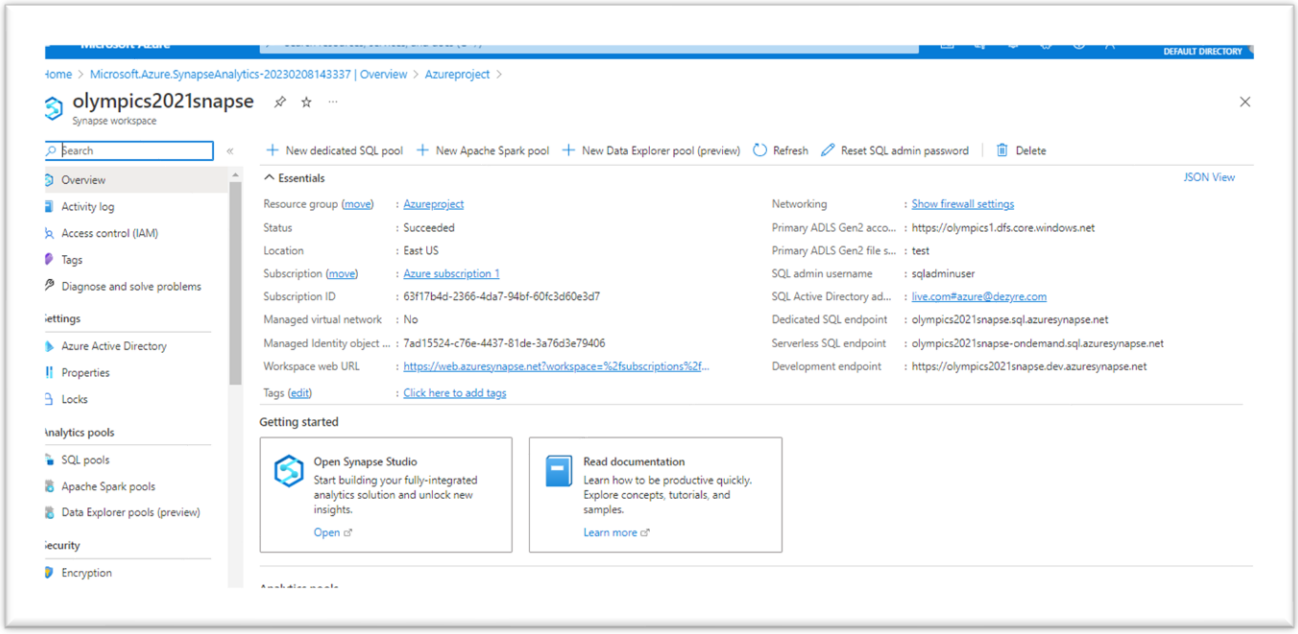
Azure Synapse Analytics is a big data and data warehousing solution offered by Microsoft as part of the Azure cloud platform. It integrates big data and data warehousing into a single service, making it easier for organizations to analyze and manage large amounts of data. Let's create an Azure Synapse Workspace.

To create an Azure Synapse Workspace, follow these steps:

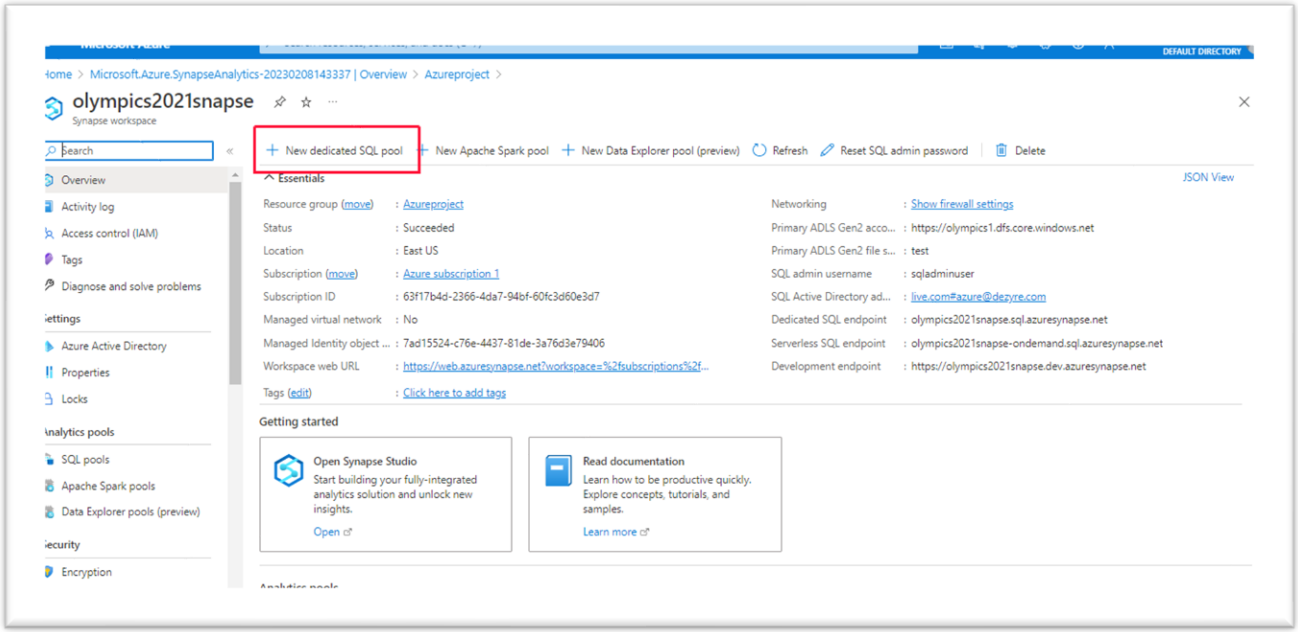
1. Log in to the Azure portal using your Azure account.
2. In the Azure portal, click on the "Create a resource" button.
3. In the search box, type "Azure Synapse Workspace" and select the result.
4. Click on the "Create" button.
5. In the "Azure Synapse Workspace" creation page, fill in the required information, such as the workspace name, subscription, resource group, and location.
6. Choose the storage account type and configure the firewall settings, if desired.
7. Click on the "Review + create" button to review the information.
8. If everything is correct, click on the "Create" button to create the Azure Synapse Workspace.

The creation process can take several minutes to complete. Once it's done, you'll have a fully functional Azure Synapse Workspace, which you can use to store, manage, and analyze big data.

Azure Synapse Workspace:



Once the Azure Synapse Workspace is created, Let's create a Dedicated SQL Pool.

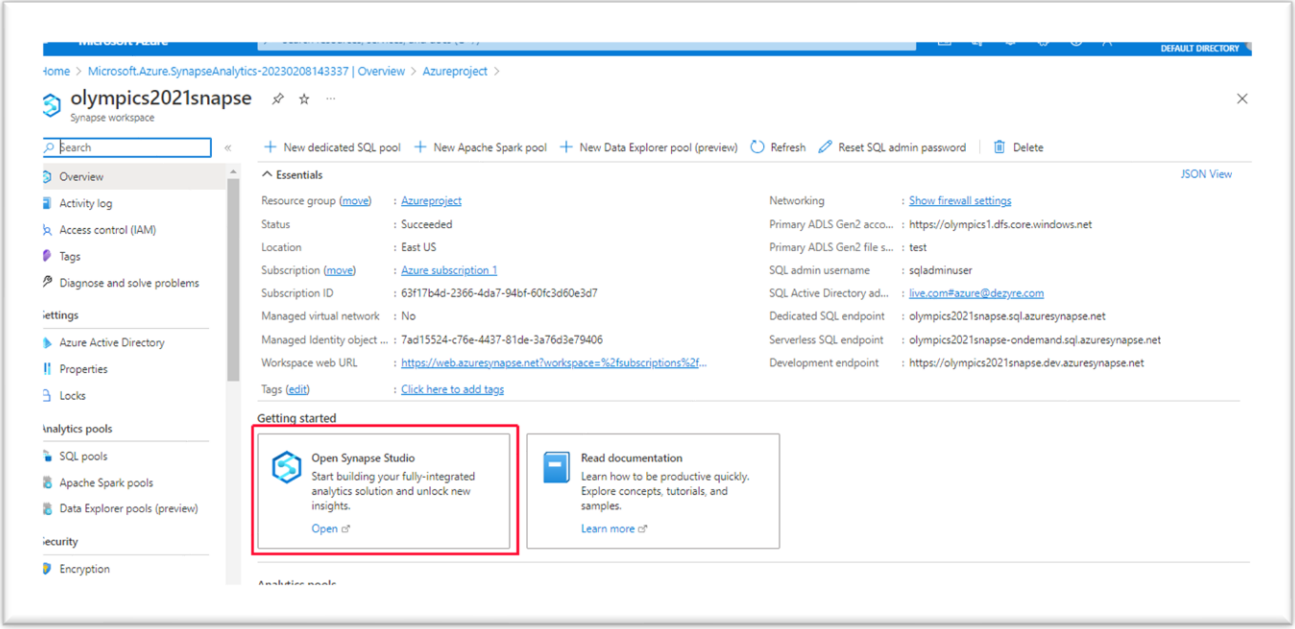


To create a dedicated SQL pool in Azure Synapse, follow these steps:

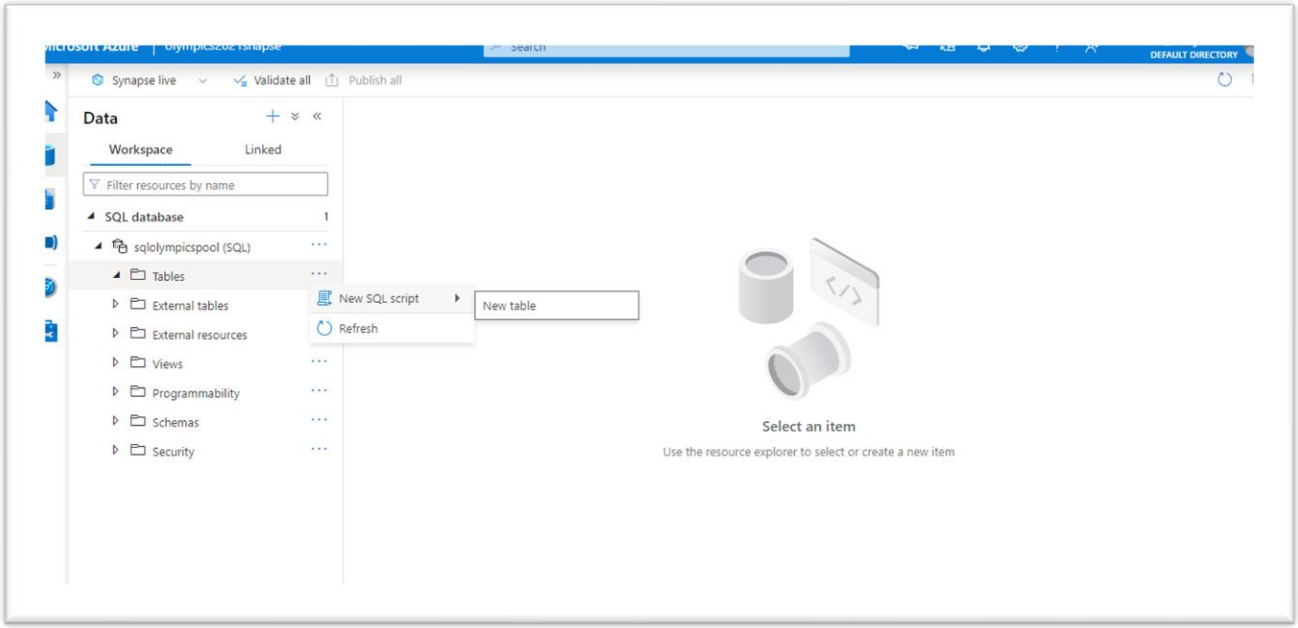
1. Log in to the Azure portal using your Azure account.
2. Find your Azure Synapse Workspace and open it.
3. In the Azure Synapse Workspace, go to the "Overview" section.
4. In the "Overview", click the "New dedicated SQL pool" button.
5. Provide a name for the SQL pool, select the subscription and resource group, and choose a region.
6. Configure the performance level and the storage capacity, and select the firewall settings.
7. Click the "Create" button to create the dedicated SQL pool.

The creation process can take several minutes to complete. Once it's done, you'll have a dedicated SQL pool that you can use to run large-scale, high-performance data warehousing workloads. You can connect to the dedicated SQL pool using a variety of tools and APIs, and start loading data, creating tables, and querying your data using T-SQL.

Let's create tables in Dedicated SQL Pools. For that, let's first open the Synapse studio.



It will open the Synapse studio in a new tab. To create tables in dedicated SQL pool, Go to Data, Workspace, SQL database, and open the newly created dedicated SQL pool. There you will find many options such as Tables, External tables, Views, etc. Click on three dots in front of the Tables, and create a new SQL script.



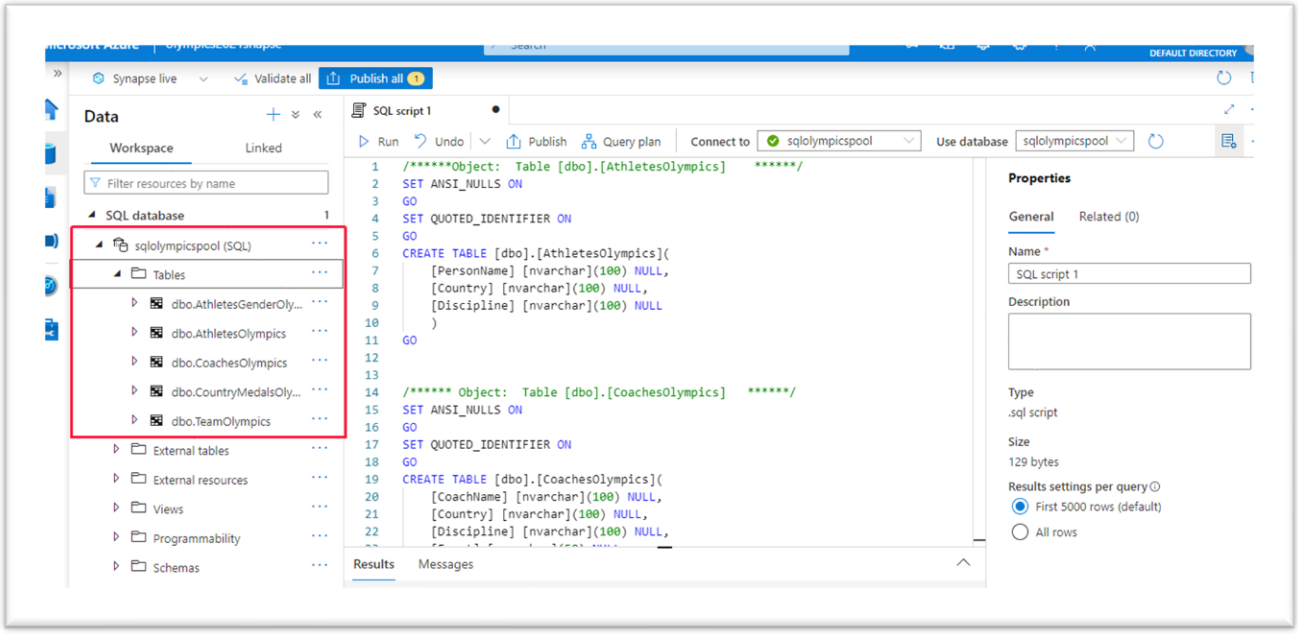
Add the following DDL scripts to create 5 tables for each data file.

DDL Script:

```
/******Object:  Table [dbo].[AthletesOlympics]    ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[AthletesOlympics](
   [PersonName] [nvarchar](100) NULL,
   [Country] [nvarchar](100) NULL,
   [Discipline] [nvarchar](100) NULL
   )
GO

/****** Object:  Table [dbo].[CoachesOlympics]   ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[CoachesOlympics](
   [CoachName] [nvarchar](100) NULL,
   [Country] [nvarchar](100) NULL,
   [Discipline] [nvarchar](100) NULL,
   [Event] [nvarchar](50) NULL
   )
GO
```

```
/****** Object:  Table [dbo].[AthletesGenderOlympics]   ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[AthletesGenderOlympics](
   [Discipline] [nvarchar](100) NULL,
   [Male] [int] NULL,
   [Female] [int] NULL,
   [TotalAthletes] [int] NULL
   )
GO

/****** Object:  Table [dbo].[CountryMedalsOlympics]   ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[CountryMedalsOlympics](
   [RankId] [int] NULL,
   [Country] [nvarchar](100) NULL,
   [Gold] [int] NULL,
   [Silver] [int] NULL,
   [Bronze] [int] NULL,
   [Total] [int] NULL,
   [RankByTotal] [int] NULL
   )
GO

/****** Object:  Table [dbo].[TeamOlympics]   ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[TeamOlympics](
   [TeamName] [nvarchar](100) NULL,
   [Country] [nvarchar](100) NULL,
   [Discipline] [nvarchar](100) NULL,
   [Event] [nvarchar](50) NULL
   )
GO
```

Once you run the above DDL script, it will create five tables as follows:



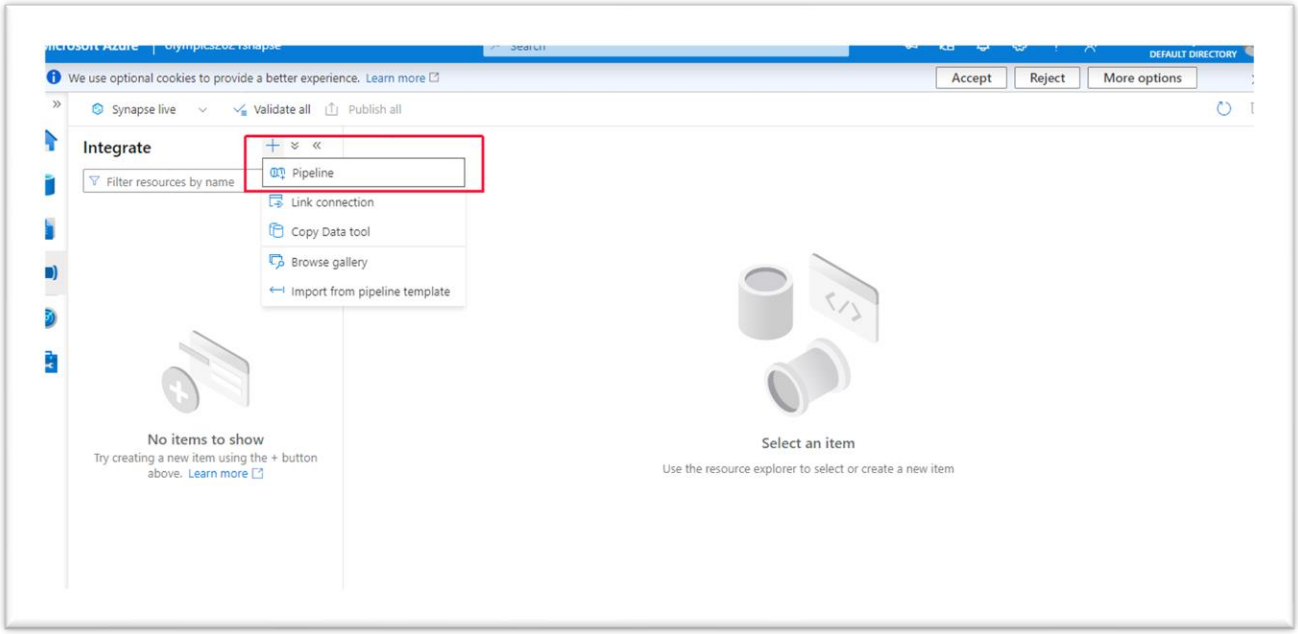
Now, we have created 5 tables in dedicated SQL Pool. Let's create an Azure Data Factory pipeline in Azure Synapse Studio to ingest data from the Azure Storage account to these 5 tables stored in a dedicated SQL Pool.

Azure Data Factory Pipeline in Synapse Studio

An Azure Data Factory (ADF) pipeline in Synapse Studio is a series of activities that perform a specific task, such as copying data or running an Azure Machine Learning model. The pipeline is created in a visual canvas within Synapse Studio, where you can select and configure activities, specify input and output connections, and set properties for each activity.
Once the pipeline is published, it can be triggered to run on a schedule or manually. The results of the pipeline can be monitored and viewed in the Synapse Studio monitor, where you can see the status of each activity and any errors that may occur.
ADF pipelines in Synapse Studio allow you to integrate data from various sources, perform transformations, and store the results in a desired location. This enables you to automate your data workflows, ensuring that data is processed accurately and efficiently.

To create a pipeline in Azure Data Factory (ADF) in Synapse Studio, follow these steps:

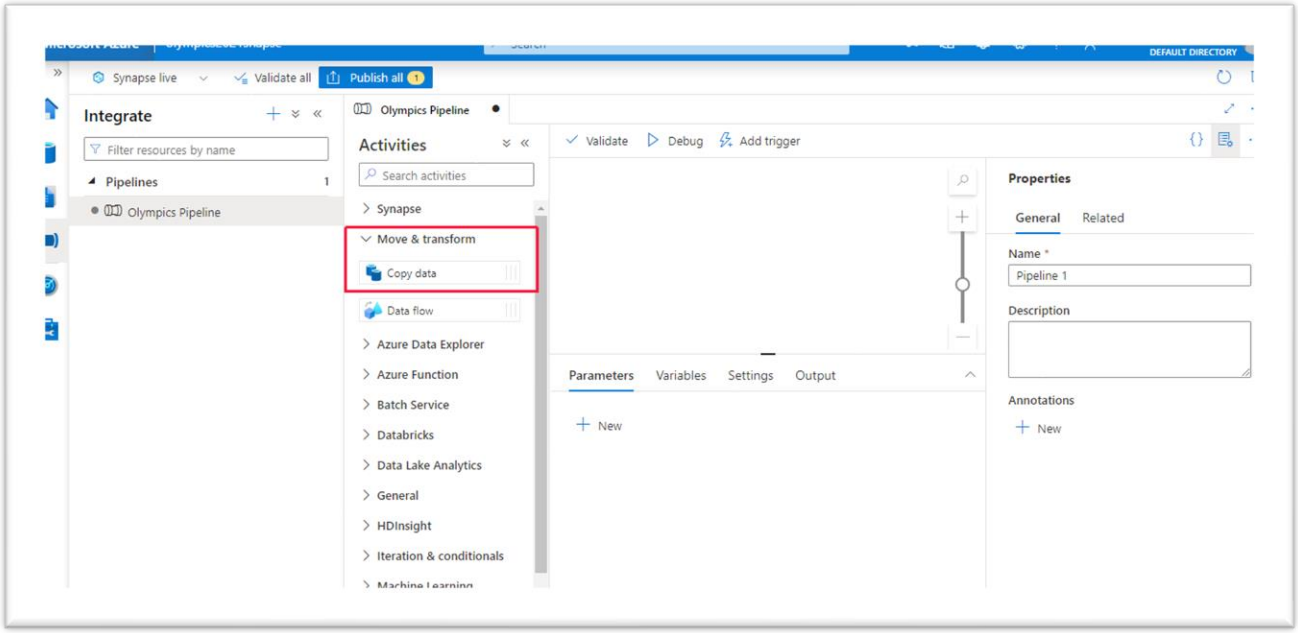
1. Go to your Synapse workspace in the Azure portal and open Synapse Studio.
2. In the left-side navigation, click "Integrate" to open ADF.
3. Click on the '+' icon and click on Pipeline.
4. It will create a new ADF pipeline in Synapse Studio



Before we start building our pipeline, let's understand an important concept called '**Linked service**' in Azure Data Factory.
A linked service in Azure Synapse is a connection to a specific data source or destination used within an Azure Data Factory (ADF) pipeline. Linked services define the authentication, connectivity, and other configuration details required to access the data source or destination.
For example, a linked service for a SQL Server database would specify the server name, database name, authentication type, and any other relevant details needed to connect to the database. A linked service for an Azure Blob storage account would specify the storage account name, access key, and other details needed to access the account.

In Azure Synapse, linked services are created and managed in the Azure portal, and then referenced within ADF pipelines to perform operations on the linked data sources and destinations. This allows you to reuse the linked services across multiple pipelines and reduces the need for manual configuration, improving consistency and ease of use.
Overall, linked services are an important part of data integration and orchestration in Azure Synapse and enable you to manage and automate your data workflows.

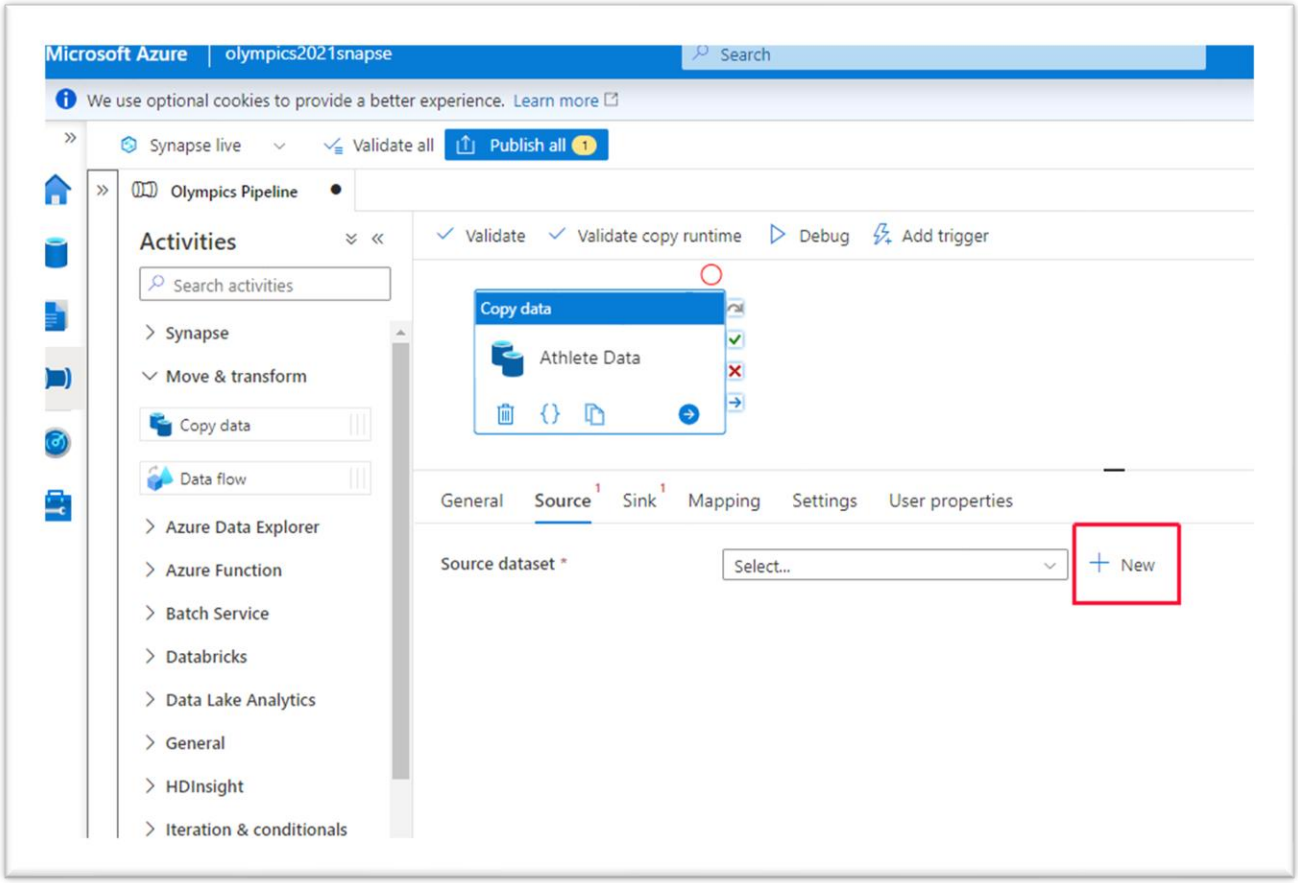
This is what the newly created pipeline looks like:



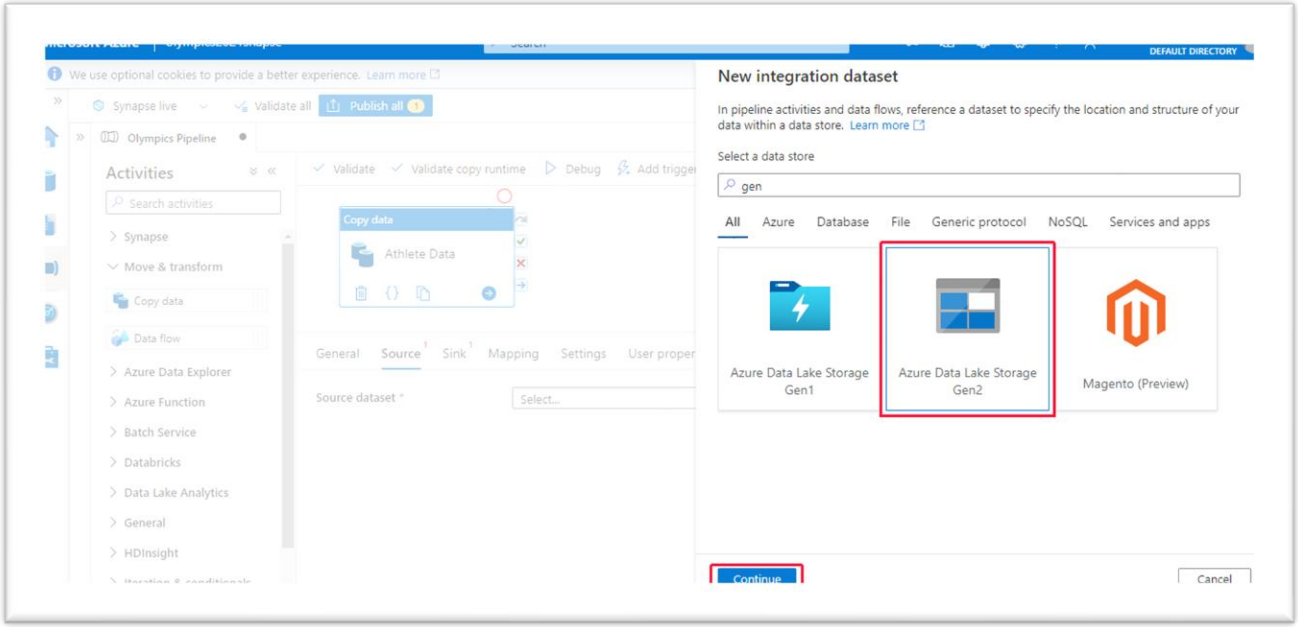
We will use the “Copy data” component. The "Copy Data" component in Azure Data Factory (ADF) allows you to move and transform data from one location to another. The component provides a visual interface for defining the source and destination of the data, as well as any transformations that need to be applied to the data during the copy process. You can use the "Copy Data" component to perform operations such as filtering, mapping, and flattening the data.
To configure the "Copy Data" component in ADF, follow these steps:

1. Drag and drop the “Copy Data” component from Move & transform to the main canvas.
2. In the "Source" section, select the data source from which you want to copy data. This could be a database, a file, or another component in the pipeline.
3. In the "Sink" section, select the data store to which you want to copy the data. This could be a database table, a file, or another component in the pipeline.

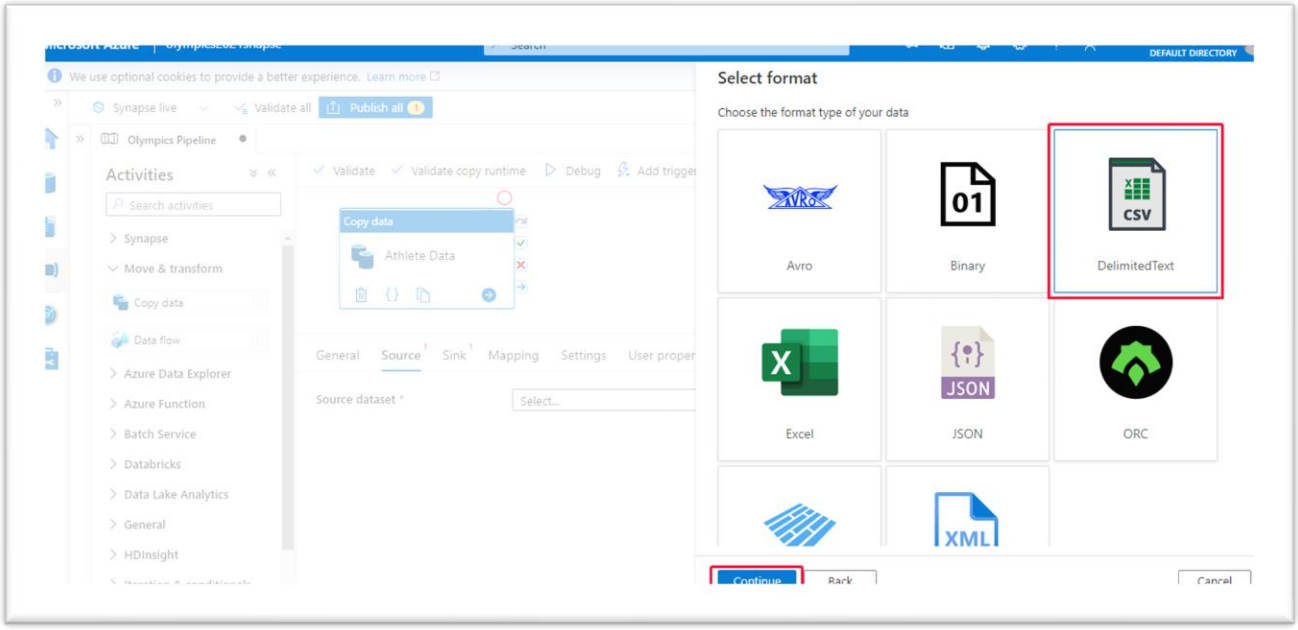
Let's configure the Source:
Click on New to create a source dataset.



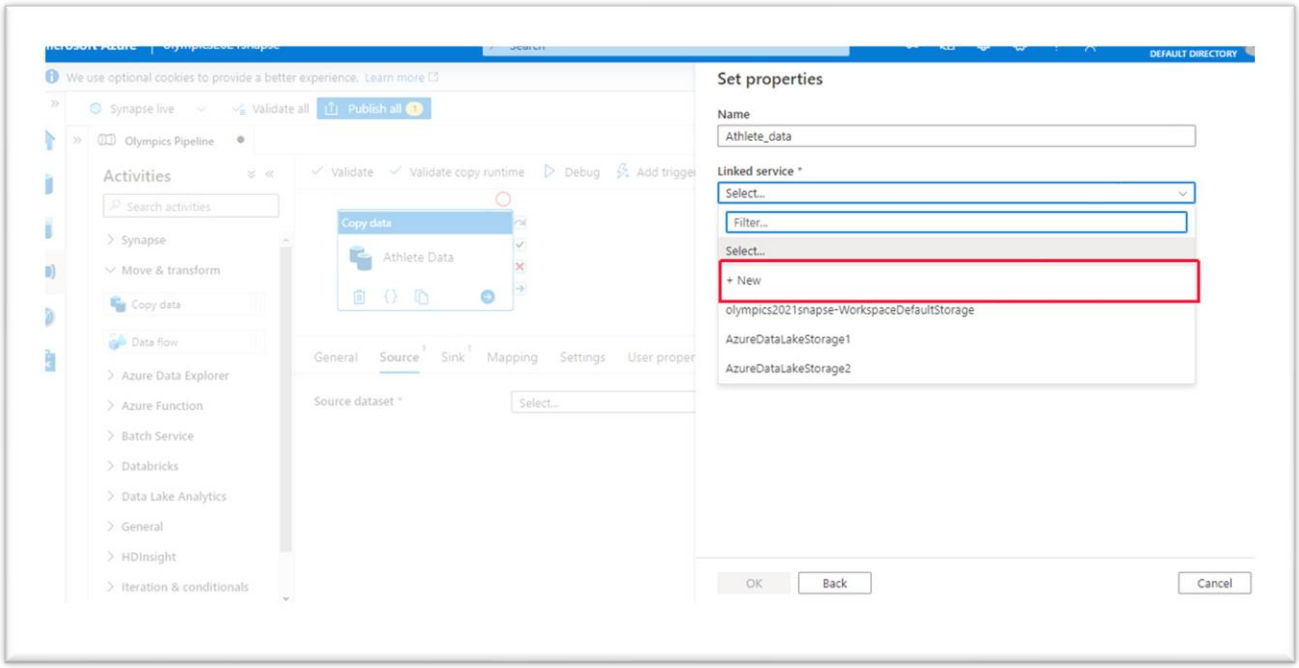
Select Azure Data Lake Storage Gen2 and click Continue.



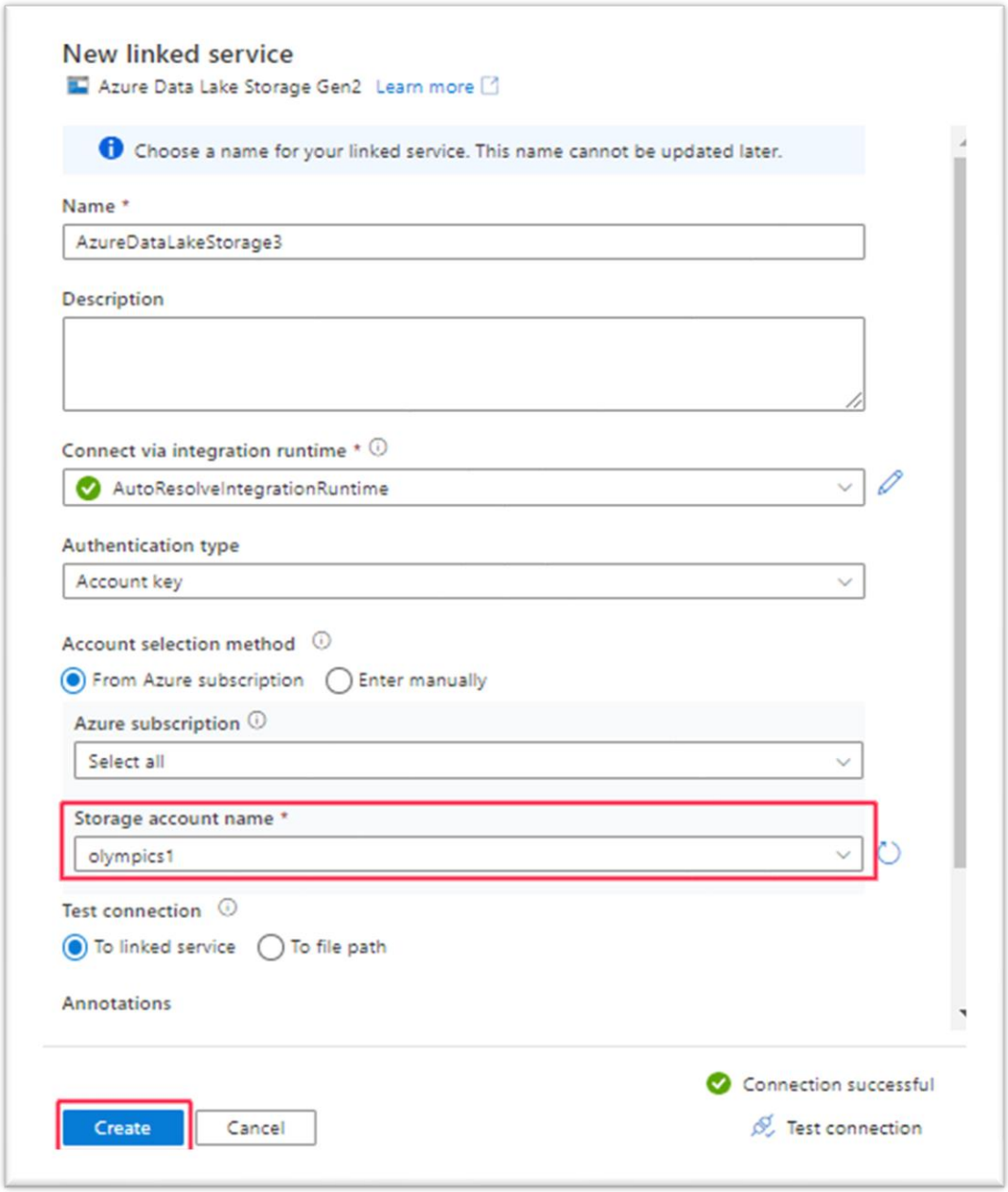
Select DelimitedText format, and click Continue.



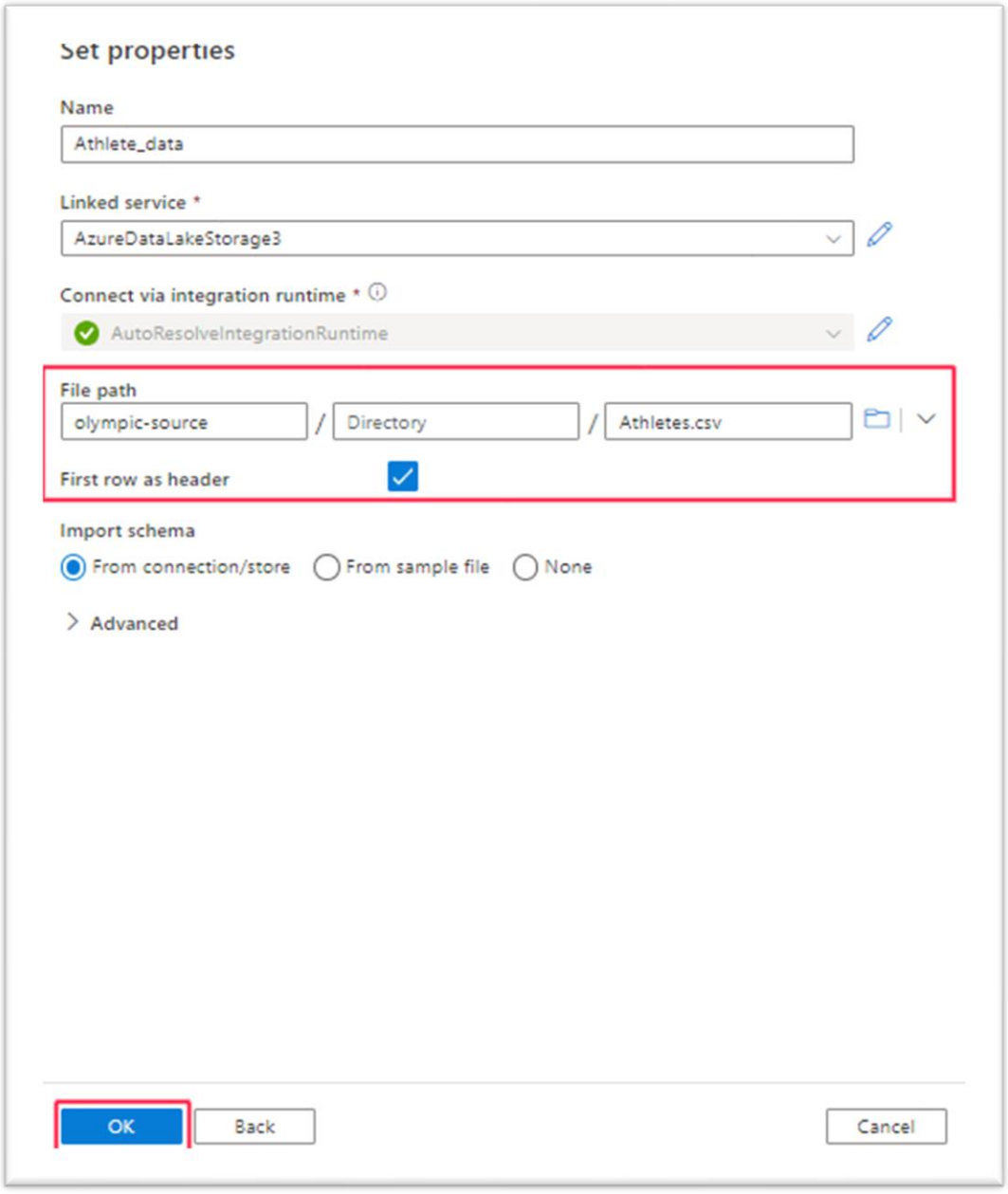
We need to create a Linked service for Azure Data Lake Storage, click on New.



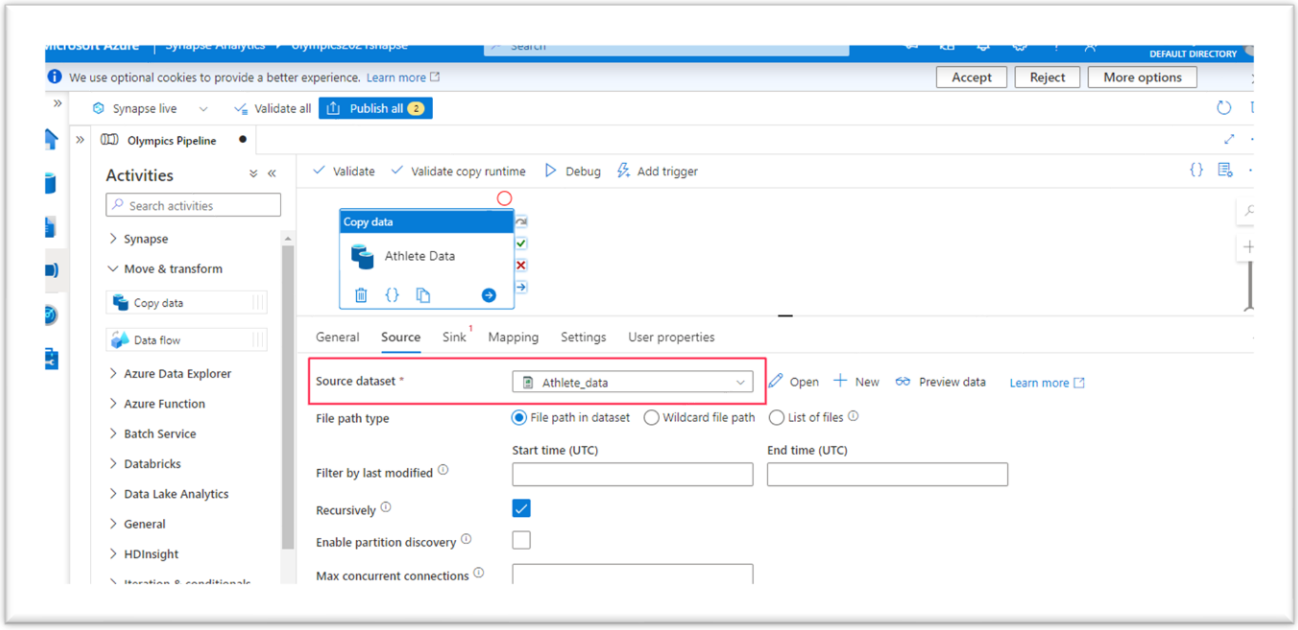
Give a name to the new linked service, please make sure you select the storage account where the CSV files are stored. Leave everything else by default, and click Create.



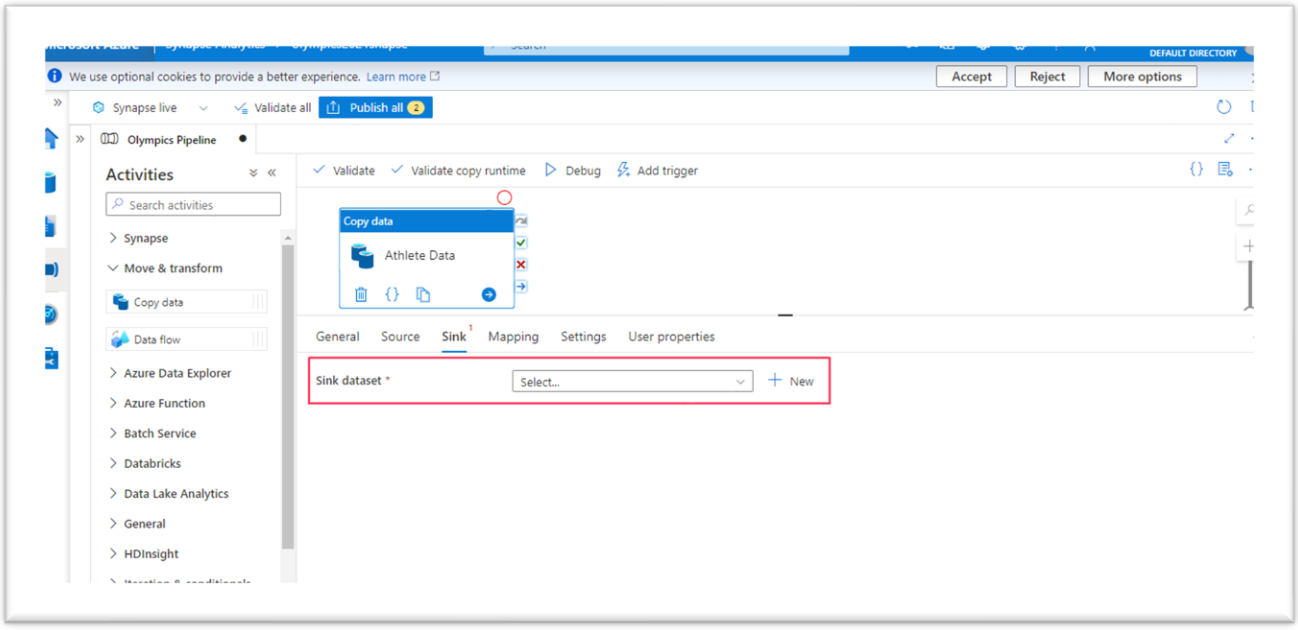
Once you create a new linked service, you will be redirected to the Set properties page. Please ensure you add the container name where the CSV files are stored here. Here the container name is the "olympic-source", and we are ingesting "Athletes.csv" file. After entering the correct information, click OK.



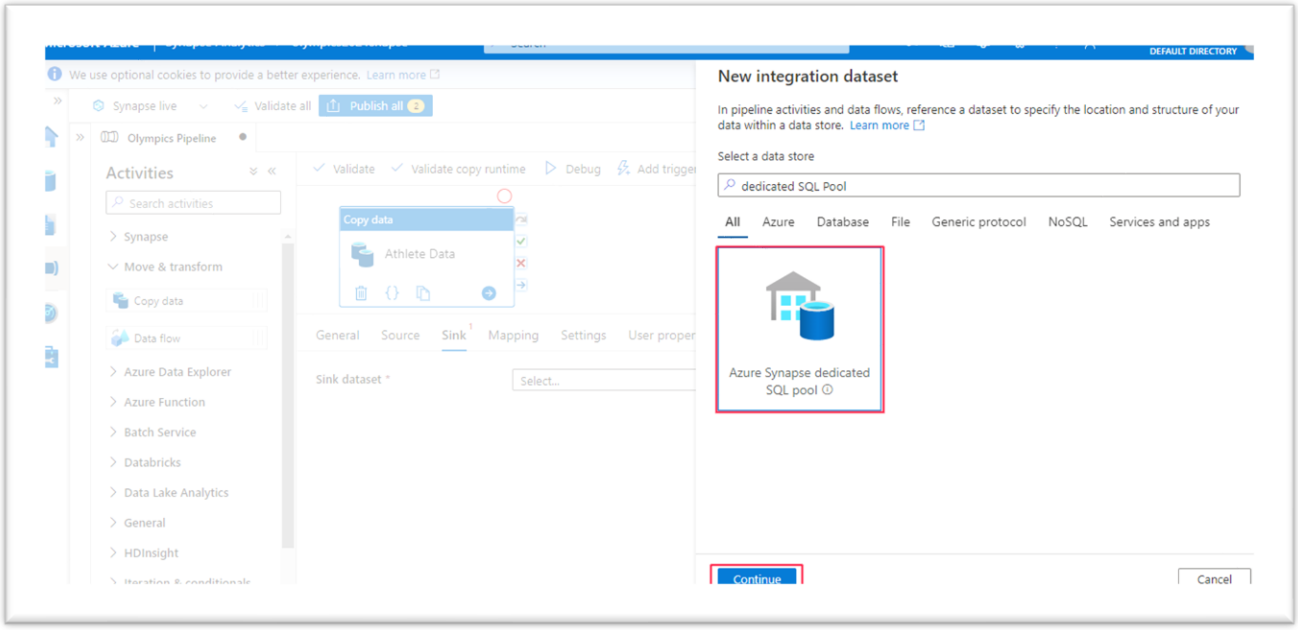
We have added the Source successfully.



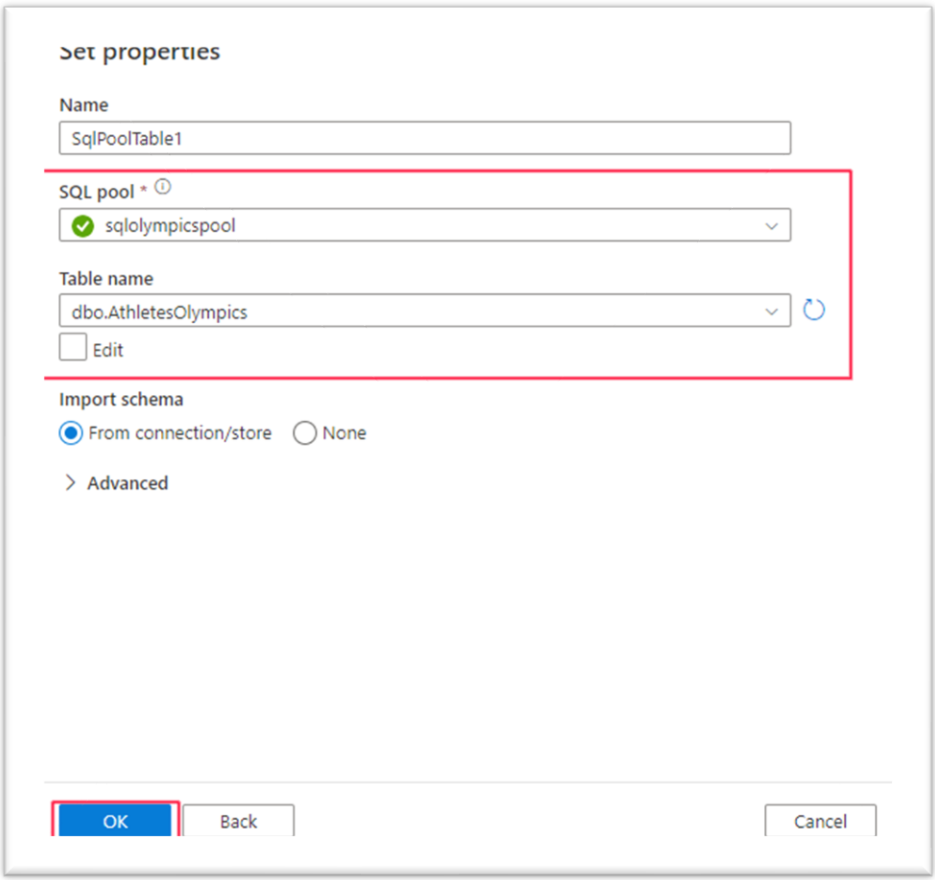
Let's configure the Sink:
Click on New to create a Sink dataset.



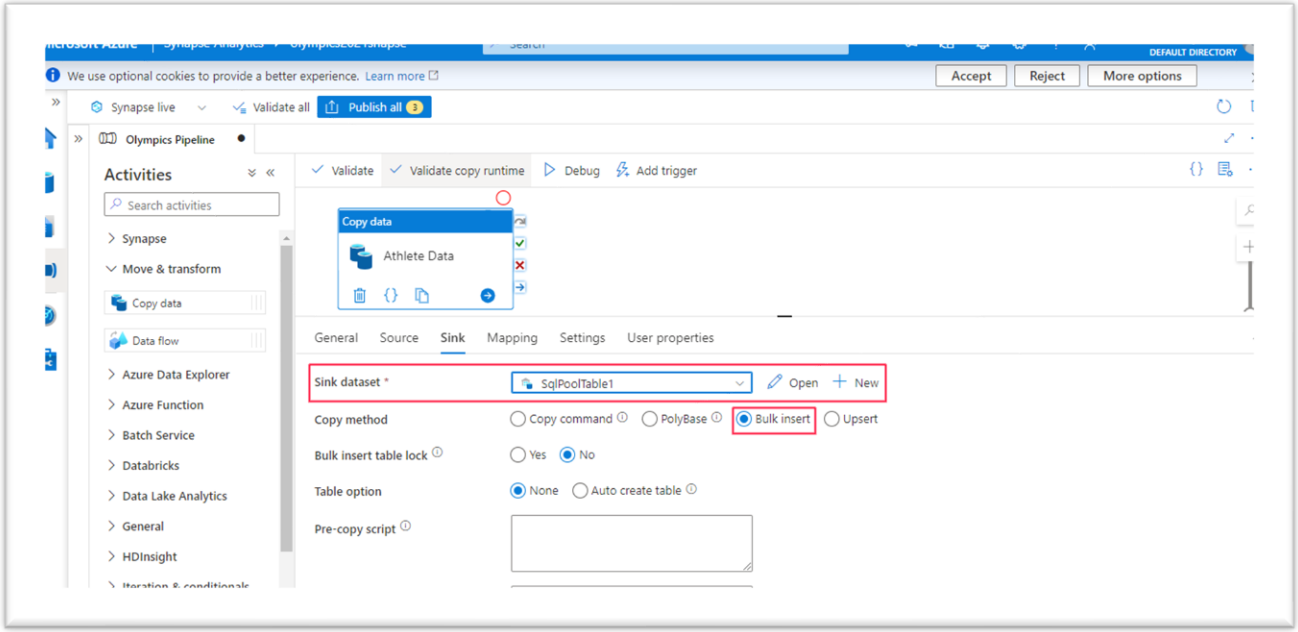
Select Azure Synapse dedicated SQL Pool and click Continue.



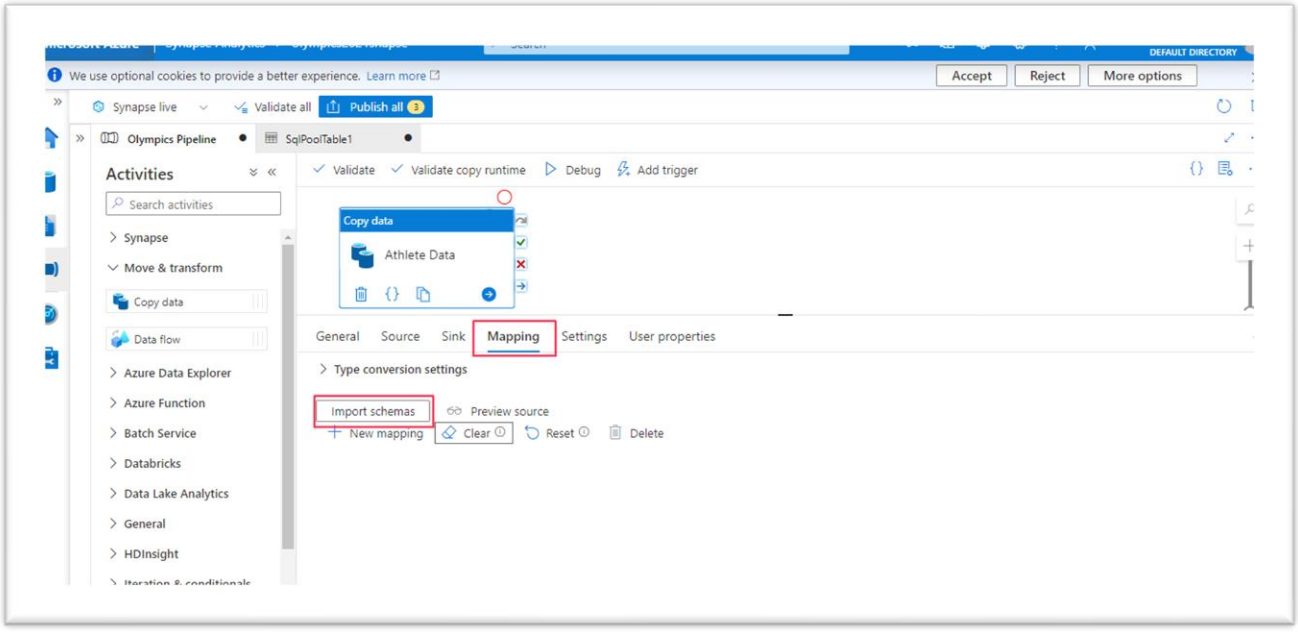
Select the SQL pool and the table where you want to ingest the data. Over here, we are trying to ingest "Athletes.csv" file, hence we select the "AthletesOlympics" table, and click OK.

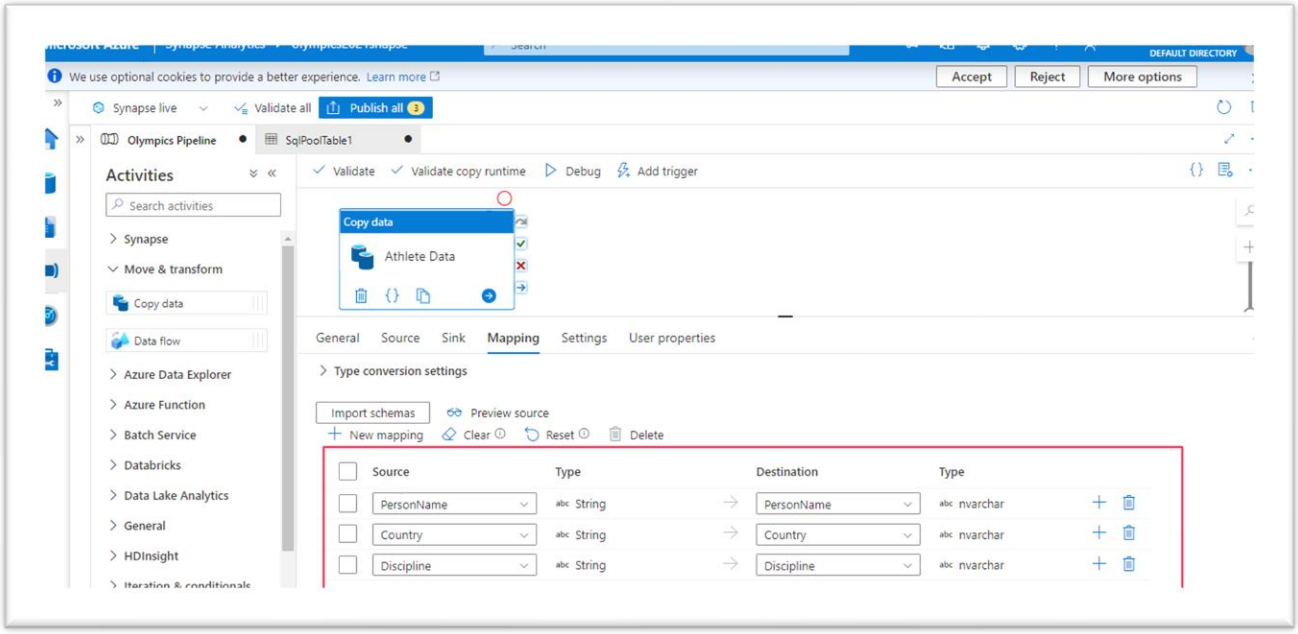


We have successfully created a Sink dataset. Make sure to select the Bulk insert copy method. The bulk insert copy method refers to a process of inserting large amounts of data into a database table using a single SQL command. This method is designed to be efficient and fast, allowing you to quickly populate a database table with large volumes of data.

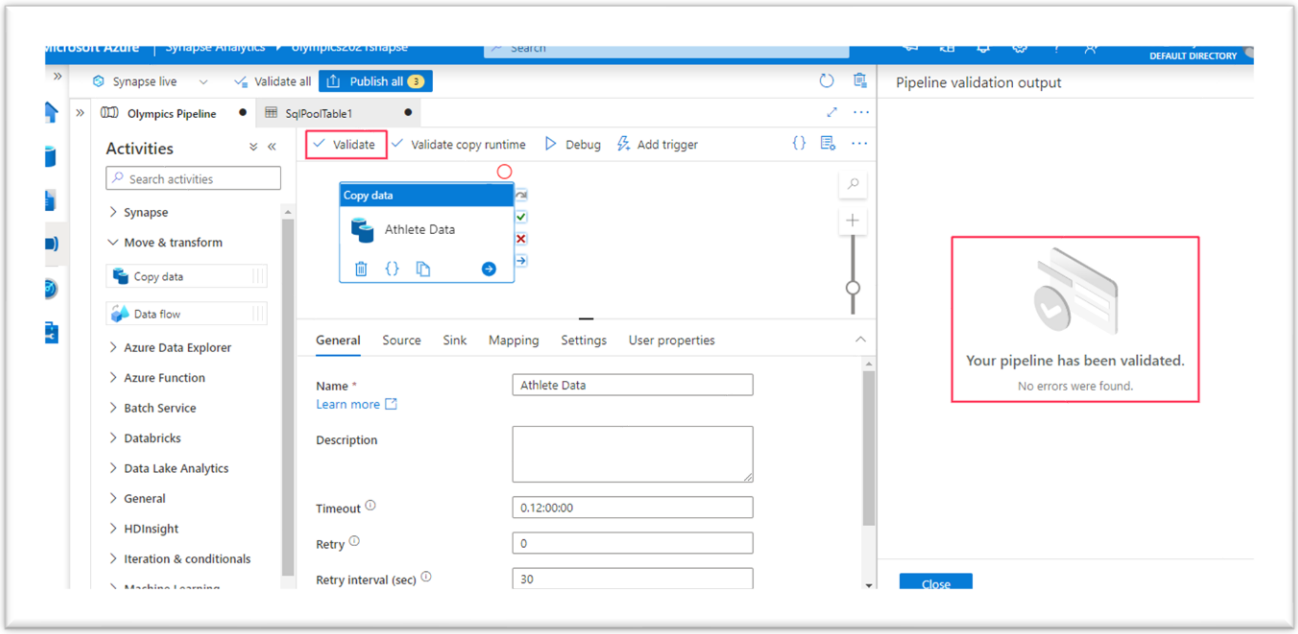


Mapping in Azure Data Factory (ADF) is the process of defining the relationships between the source and target data in a data integration scenario. This is achieved by using mapping data flows, which are a visual, no-code interface for transforming and shaping data. Click on Import schemas, it will automatically map the columns of the source dataset and sink dataset.

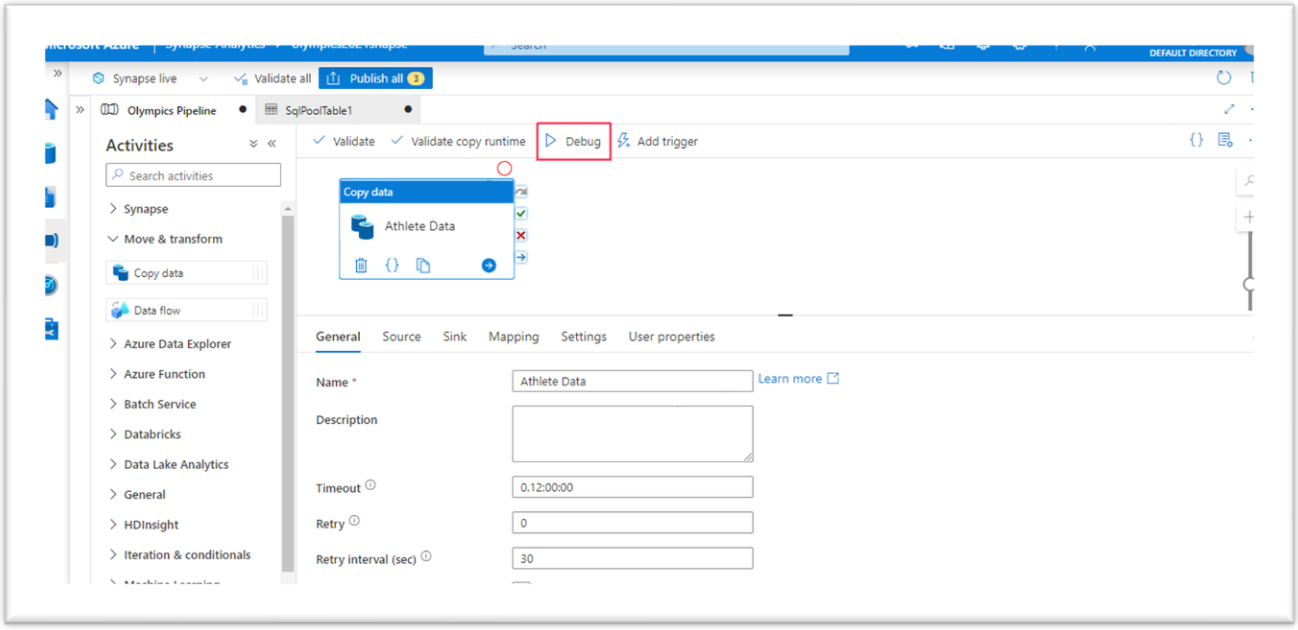




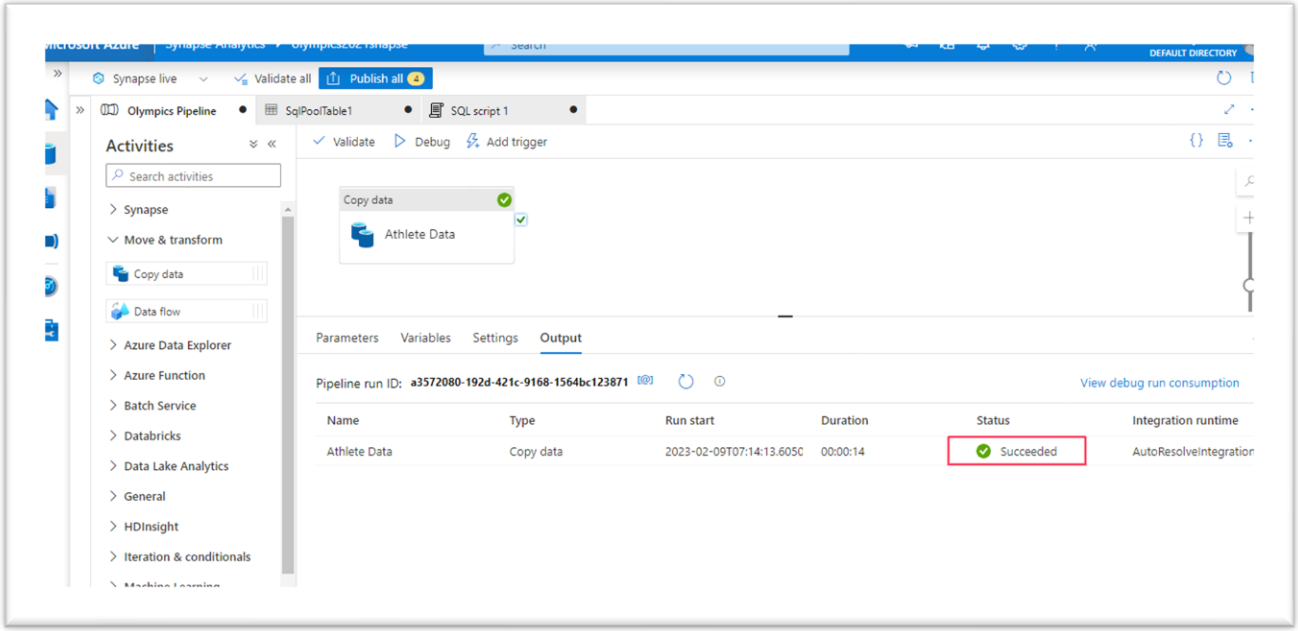
Let's validate our pipeline to check if there are any errors in the pipeline. We can see there are no errors in our pipeline, so we can proceed with running the pipeline.



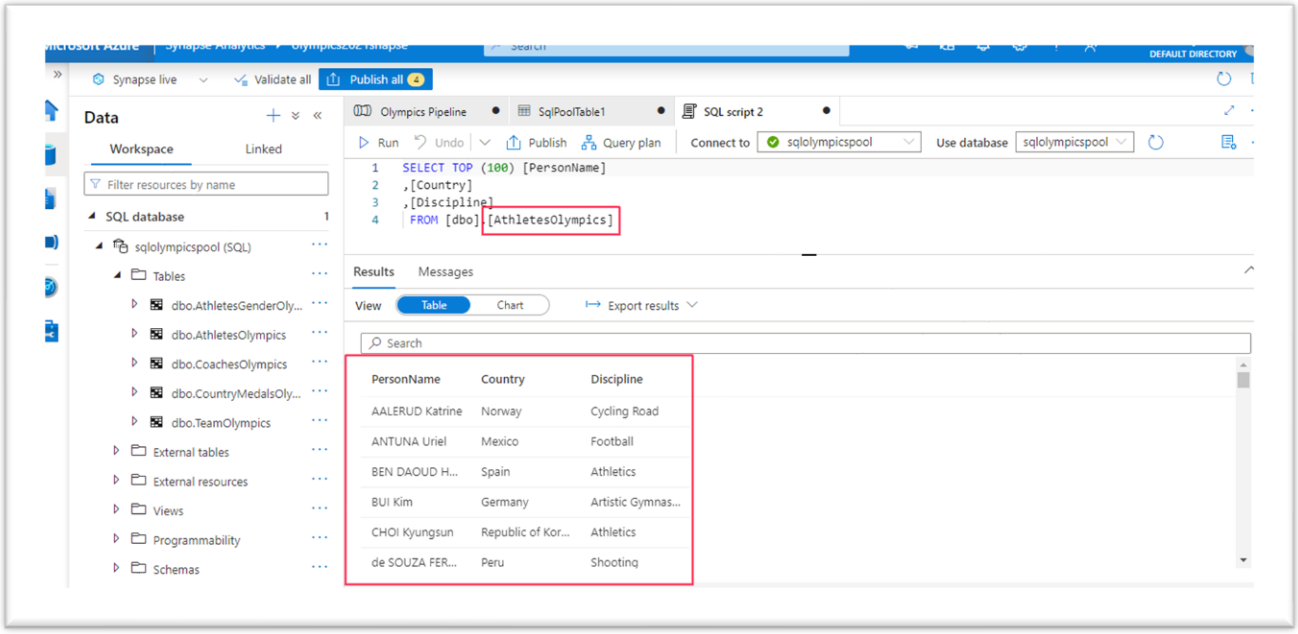
To run the pipeline, click on Debug.



The execution of the pipeline is done. We can see the status as 'Succeeded'.

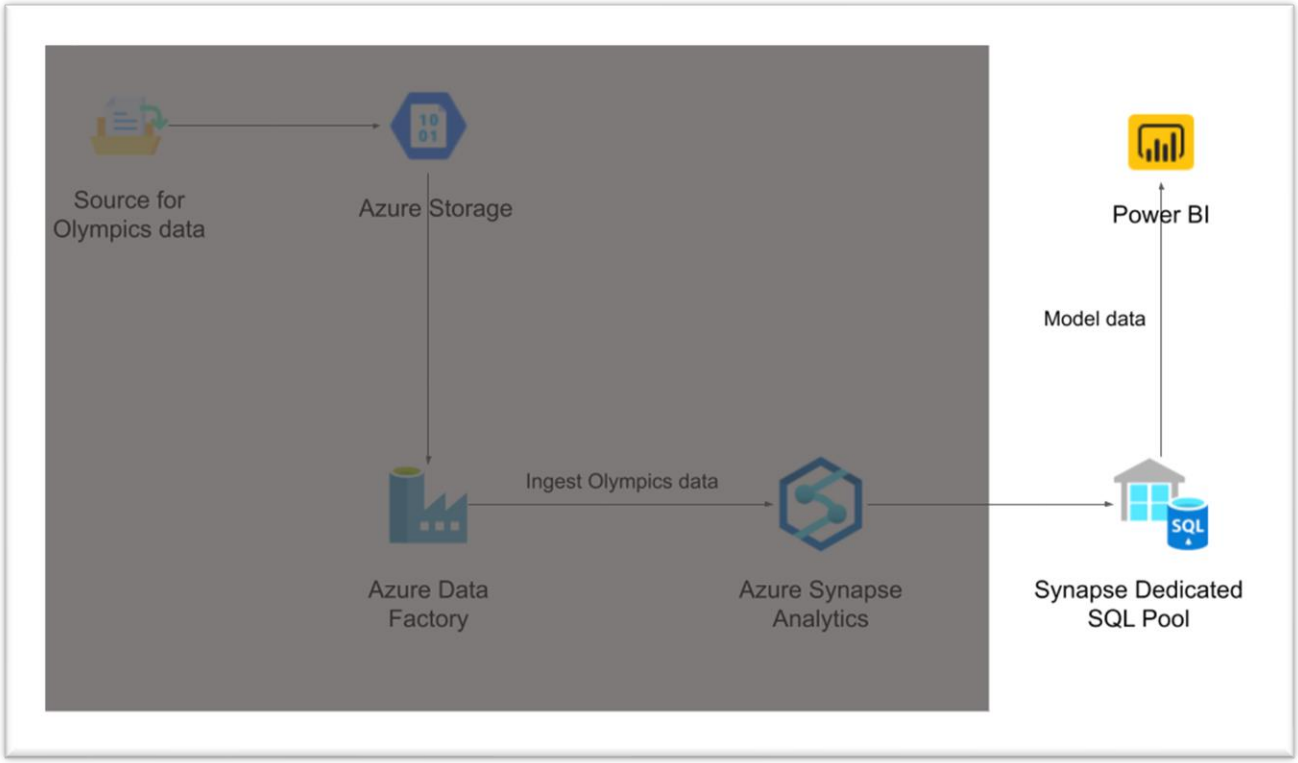


Let's check the 'AthletesOlympics' table.
We can see the data is present in the 'AthletesOlympics' table.



We have successfully copied the data of 'Athletes.csv' file stored in Azure storage into the 'AthletesOlympics' table stored in dedicated SQL Pool. Please follow the same steps for the remaining data files.

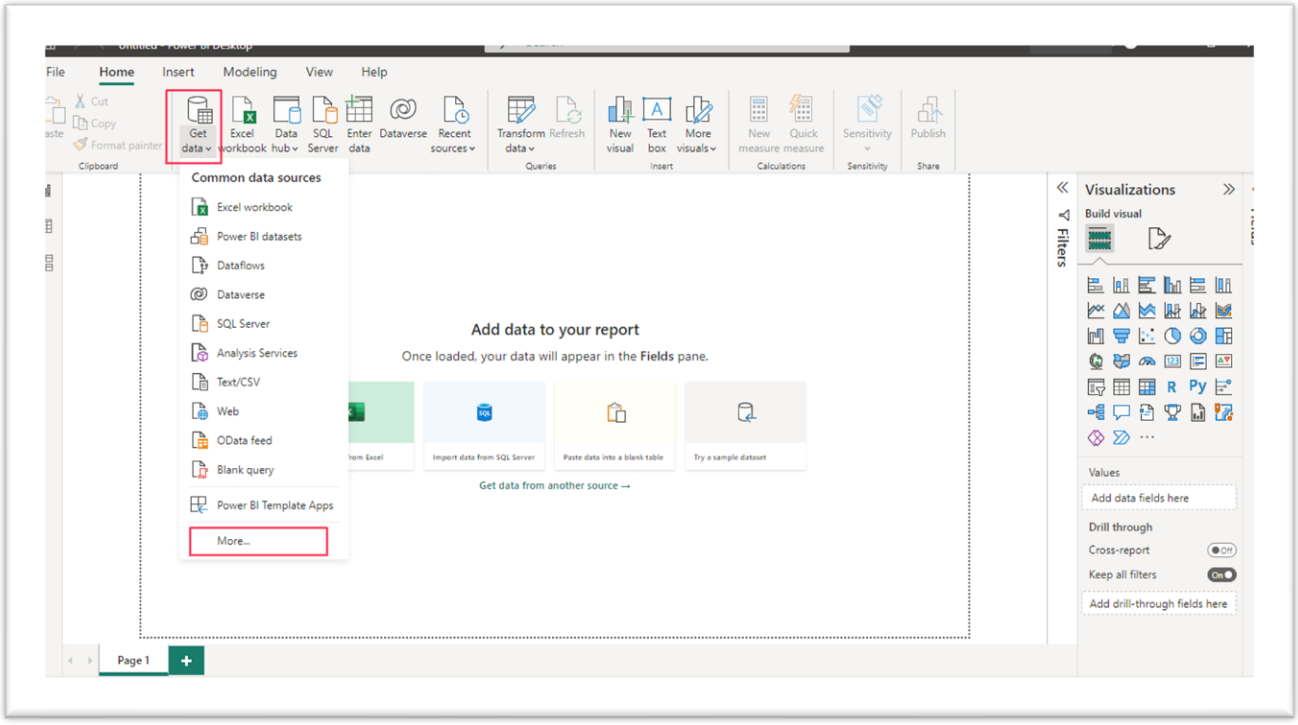
Data Visualization in Power BI



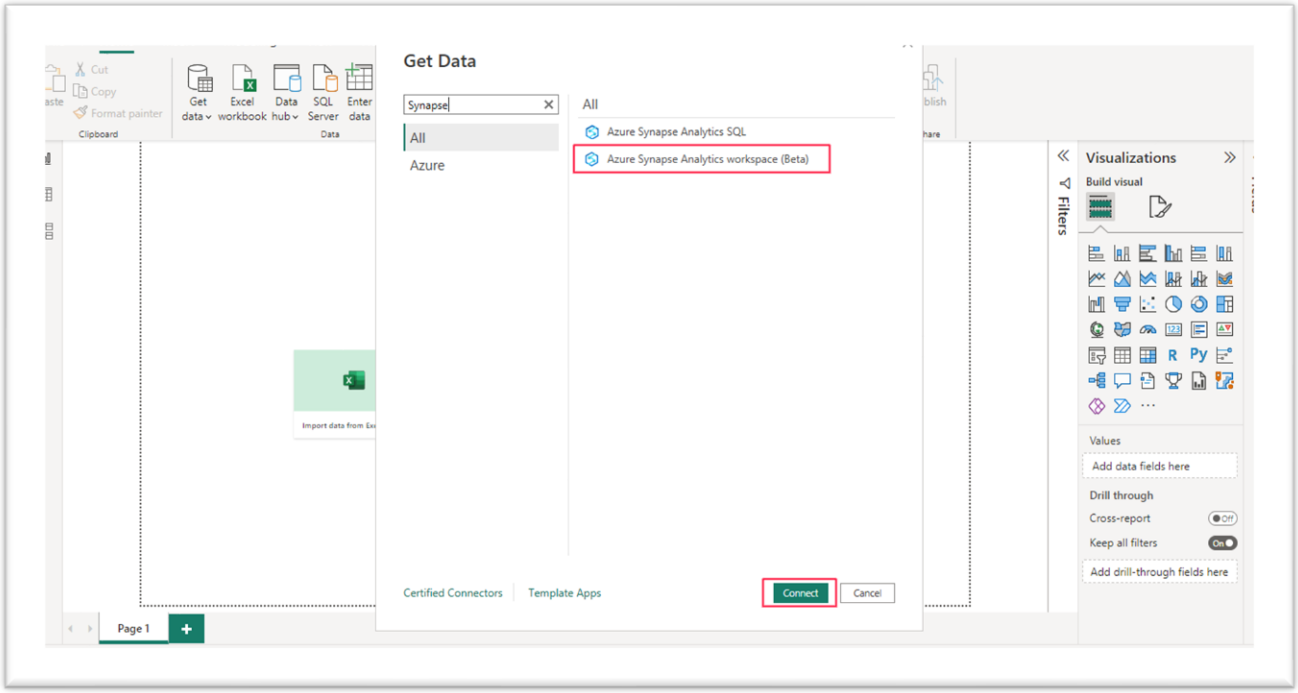
Power BI is a suite of business analytics tools from Microsoft that allows you to visualize, analyze, and share data in meaningful ways. It provides a wide range of features and capabilities, including interactive visualizations, data exploration, and data sharing. With Power BI, you can connect to a variety of data sources, including on-premises databases, cloud-based data stores, and online services like Excel, Salesforce, and Google Analytics. You can then use the intuitive Power BI interface to build reports and dashboards that display your data in interactive charts, tables, and visuals.
Power BI offers a number of features that make it easy to work with your data, including:

- Data connectivity: You can connect to a wide range of data sources, including SQL Server, Excel, SharePoint, and more.
- Data modeling: Power BI includes a powerful data modeling engine that allows you to define relationships, calculated columns, and measure expressions.
- Data visualization: With Power BI, you can create interactive visuals, such as charts, tables, and maps, that display your data in meaningful ways.
- Report sharing: You can publish and share your reports with others, either by embedding them in other applications or by sharing them through the Power BI service.

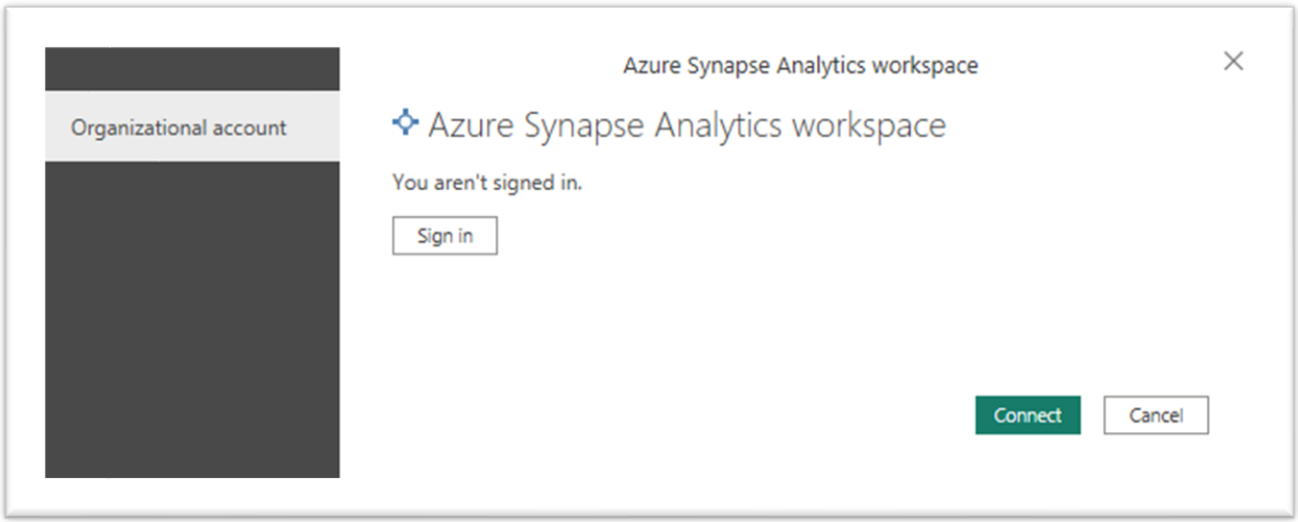
Let's connect to Azure Synapse Analytics Workspace. Click on Get Data -> More.



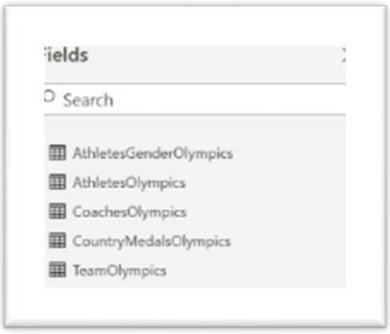
Search for Azure Synapse Analytics workspace (Beta), and click on connect.



Provide the credentials. Select the dedicated SQL Pool and tables to use in Power BI for visualization.



The selected tables will appear in the Fields section of Power BI.



Let's create visualizations in Power BI.

Summary

- In this training we began with the notion of Traditional data vs Big data and discussed the advantages of Azure Synapse Analytics over other services.
- We understood the overall architecture of the project along with the data description.
- Various components of Azure Synapse Analytics were discussed.
- We saw the different pools available in Azure Synapse Analytics.
- We observed the key differences between SQL Pool and Spark Pool.
- We created an Azure Storage account and uploaded CSV files in the container.
- We created an Azure Synapse Workspace and dedicated SQL Pool.
- We created a pipeline in Synapse studio to ingest data from Azure Storage to the dedicated SQL pool tables.
- Performed visualizations on the data using Power BI.