Explore Spark Streaming in Azure Synapse Analytics

In this exercise, you'll use *Spark Structured Streaming* and *delta tables* in Azure Synapse Analytics to process streaming data.

This lab will take approximately **15** minutes to complete.

Before you start

You'll need an [Azure subscription](#) in which you have administrative-level access.

Provision a Synapse Analytics workspace

To use Synapse Analytics, you must provision a Synapse Analytics Workspace resource in your Azure subscription.

1. Open the Azure portal at [Azure portal](#), and sign in using the credentials associated with your Azure subscription.
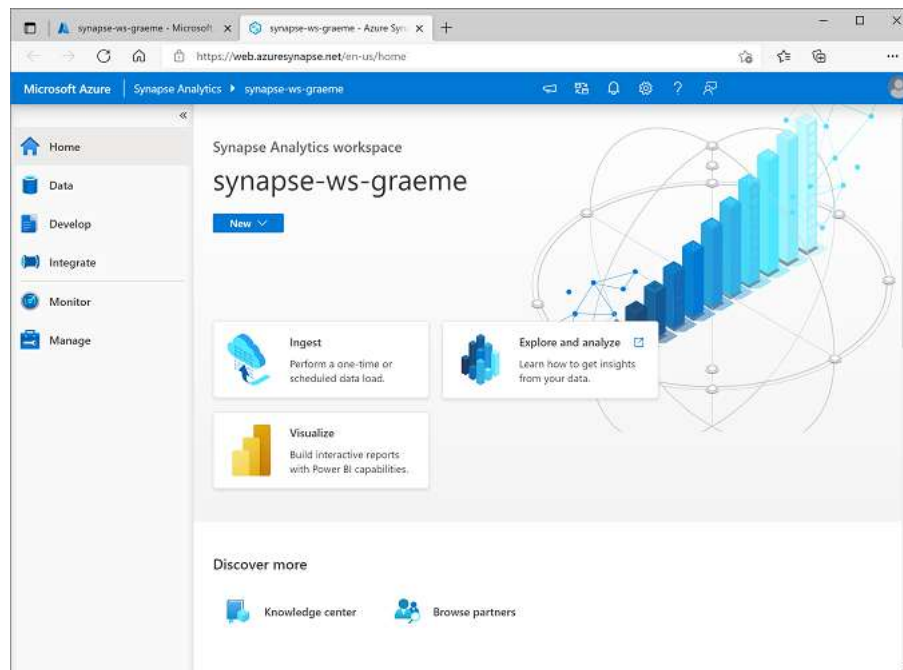
   > ⓘ **Note**: Ensure you are working in the directory containing your own subscription - indicated at the top right under your user ID. If not, select the user icon and switch directory.

2. In the Azure portal, on the **Home** page, use the **+ Create a resource** icon to create a new resource.
3. Search for *Azure Synapse Analytics*, and create a new **Azure Synapse Analytics** resource with the following settings:

   - **Subscription**: *Your Azure subscription*
       - **Resource group**: *Create a new resource group with a suitable name, like "synapse-rg"*
       - **Managed resource group**: *Enter an appropriate name, for example "synapse-managed-rg".*
   - **Workspace name**: *Enter a unique workspace name, for example "synapse-ws-*.*
   - **Region**: *Select any available region*.
   - **Select Data Lake Storage Gen 2**: From subscription
       - **Account name**: *Create a new account with a unique name, for example "datalake"*.
       - **File system name**: *Create a new file system with a unique name, for example "fs"*.

   > ⓘ **Note**: A Synapse Analytics workspace requires two resource groups in your Azure subscription; one for resources you explicitly create, and another for managed resources used by the service. It also requires a Data Lake storage account in which to store data, scripts, and other artifacts.

4. When you've entered these details, select **Review + create**, and then select **Create** to create the workspace.
5. Wait for the workspace to be created - this may take five minutes or so.
6. When deployment is complete, go to the resource group that was created and notice that it contains your Synapse Analytics workspace and a Data Lake storage account.
7. Select your Synapse workspace, and in its **Overview** page, in **Open Synapse Studio** card, select **Open** to open Synapse Studio in a new browser tab. Synapse Studio is a web-based interface that you can use to work with your Synapse Analytics workspace.

8. On the left side of Synapse Studio, use the ›› icon to expand the menu - this reveals the different pages within Synapse Studio that you will use to manage resources and perform data analytics tasks, as shown here:

Create a Spark pool

To use Spark to process streaming data, you need to add a Spark pool to your Azure Synapse workspace.

1. In Synapse Studio, select the **Manage** page.
2. Select the **Apache Spark pools** tab, and then use the **+ New** icon to create a new Spark pool with the following settings:

   - **Apache Spark pool name**: sparkpool
   - **Node size family**: Memory Optimized
   - **Node size**: Small (4 vCores / 32 GB)
   - **Autoscale**: Enabled
   - **Number of nodes** 3—-3
3. Review and create the Spark pool, and then wait for it to be deployed (which may take a few minutes).

Explore stream processing

To explore stream processing with Spark, you'll use a notebook that contains Python code and notes to help you perform some basic stream processing with Spark Structured Streaming and delta tables.

1. Download the Structured Streaming and Delta Tables.ipynb notebook to your local computer (if the notebook is opened as a text file in your browser, save it to a local folder; being careful to save it as **Structured Streaming and Delta Tables.ipynb**, not as a .txt file)
2. In Synapse Studio, select the **Develop** page.
3. On the **+** menu, select ↩ **Import**, and select the **Structured Streaming and Delta Tables.ipynb** file on your local computer.
4. Follow the instructions in the notebook to attach it to your Spark pool and run the code cells it contains to explore various ways to use Spark for stream processing.

Delete Azure resources

> ⚠ **Note**: If you intend to complete other exercises that use Azure Synapse Analytics, you can skip this section. Otherwise, follow the steps below to avoid unnecessary Azure costs.

1. Close the Synapse Studio browser tab, without saving any changes, and return to the Azure portal.
2. On the Azure portal, on the **Home** page, select **Resource groups**.

3. Select the resource group for your Synapse Analytics workspace (not the managed resource group), and verify that it contains the Synapse workspace, storage account, and Data Explorer pool for your workspace (if you completed the previous exercise, it will also contain a Spark pool).
4. At the top of the **Overview** page for your resource group, select **Delete resource group**.

5. Enter the resource group name to confirm you want to delete it, and select **Delete**.

   After a few minutes, your Azure Synapse workspace and the managed workspace associated with it will be deleted.