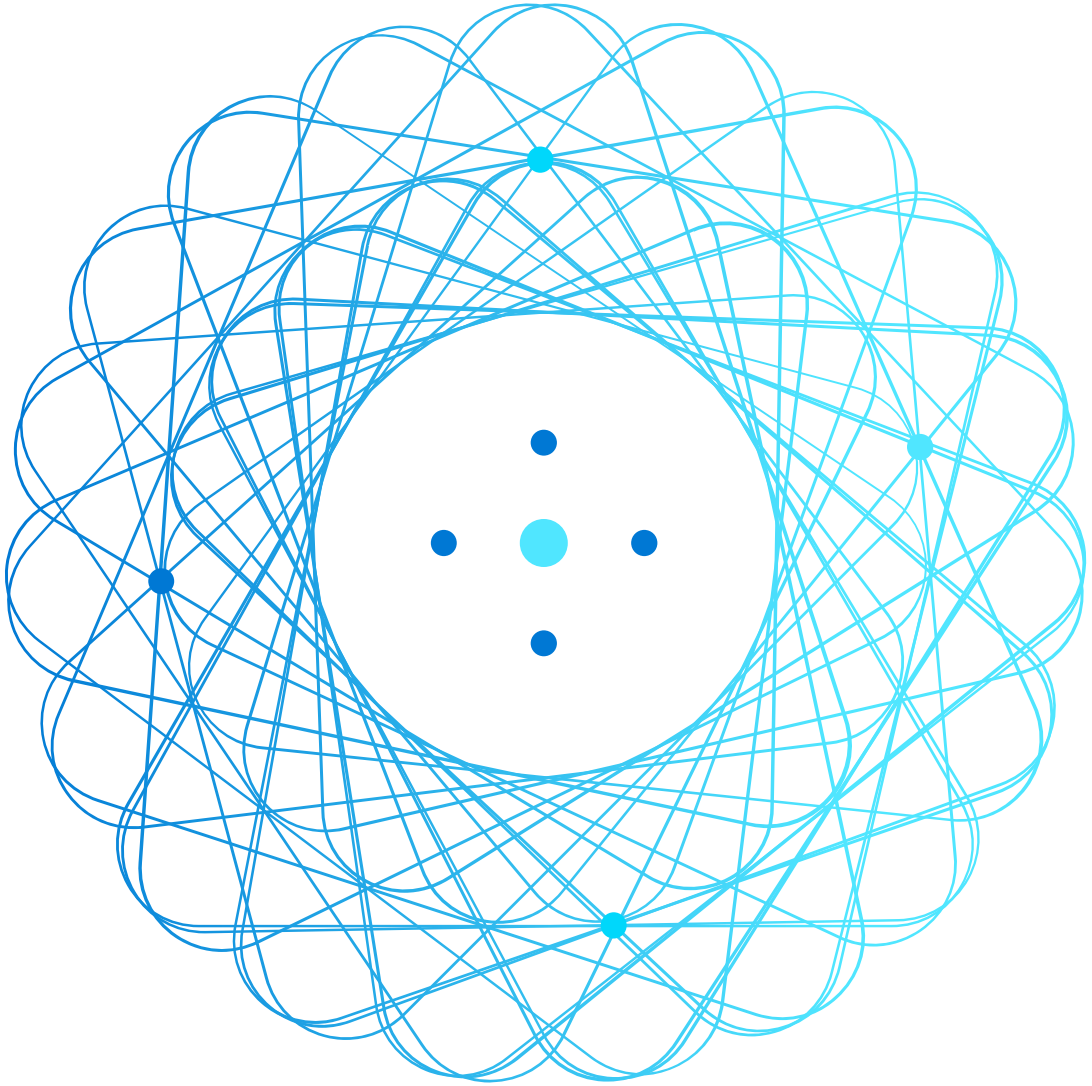
Microsoft Azure

Transfer and transform data with Azure Synapse Analytics Pipelines



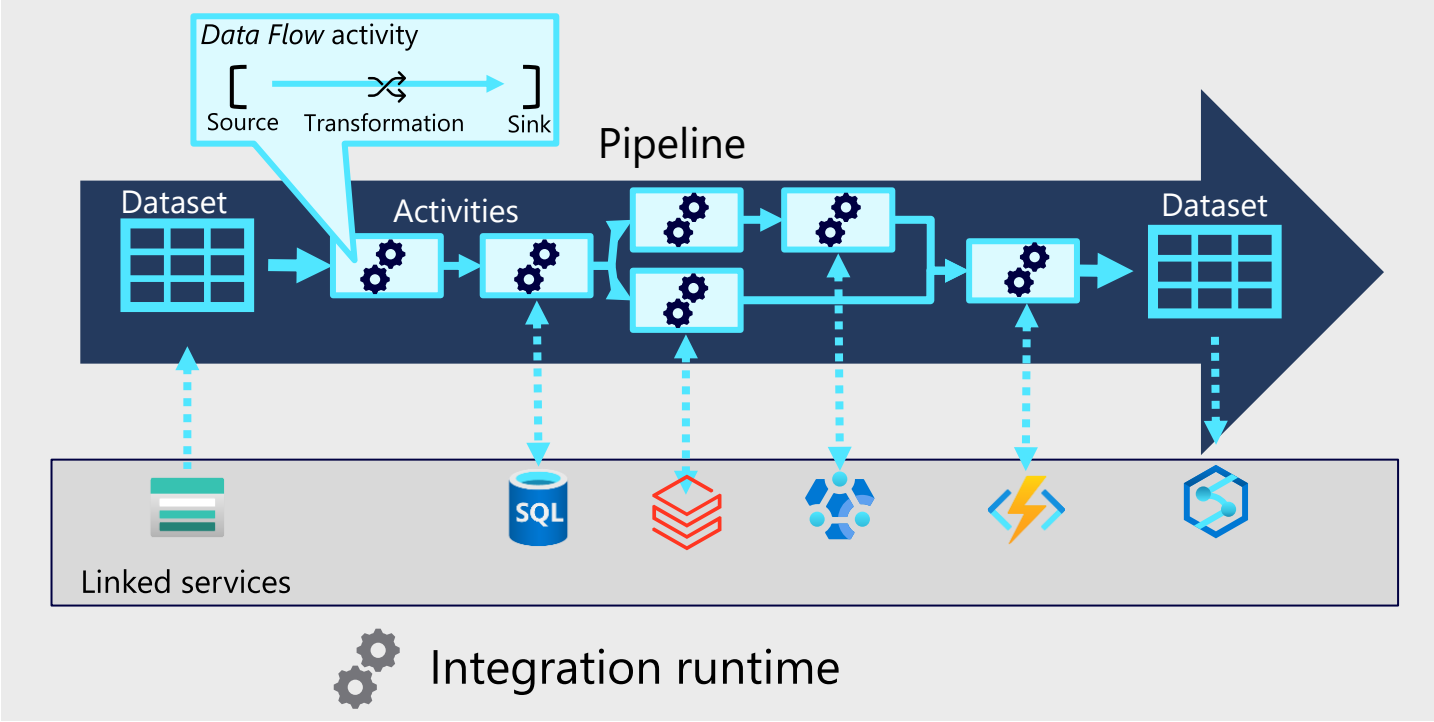
Agenda

- Build a data pipeline in Azure Synapse Analytics
- Use Spark Notebooks in an Azure Synapse Pipeline

Build a data pipeline in Azure Synapse Analytics

Understand pipelines

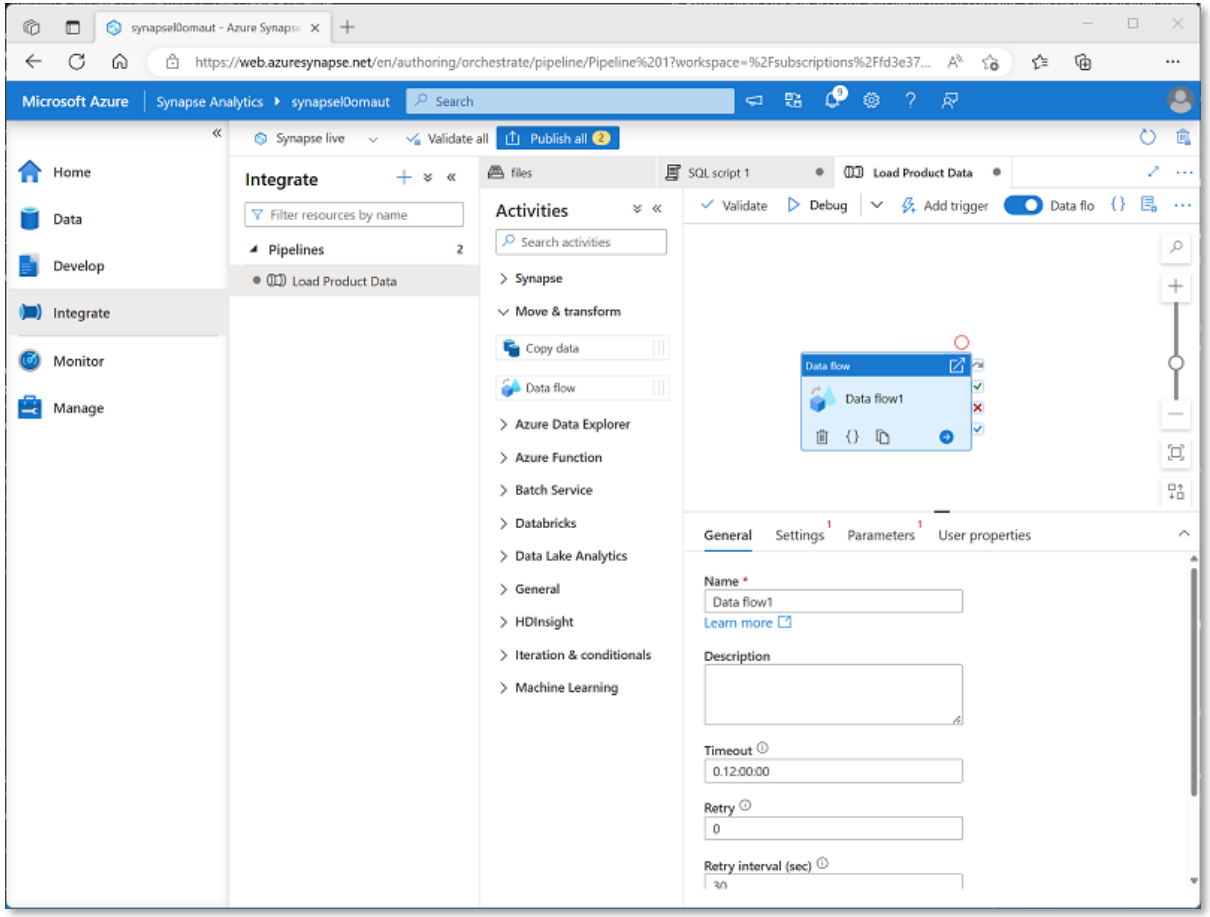
- Pipelines encapsulate a flow of *activities* that are orchestrated by an *integration runtime*
- Activities can include:
  - *Data movement* and *data transformation* activities that transfer data from *sources* to *sinks*
  - External processing activities
  - *Control flow* activities that manage variables and processing logic
- *Linked services* provide access to data stores and processing platforms where activities can be run
- The data processed in a pipeline is defined in *datasets*, accessed through linked services



Create a pipeline in Azure Synapse Studio

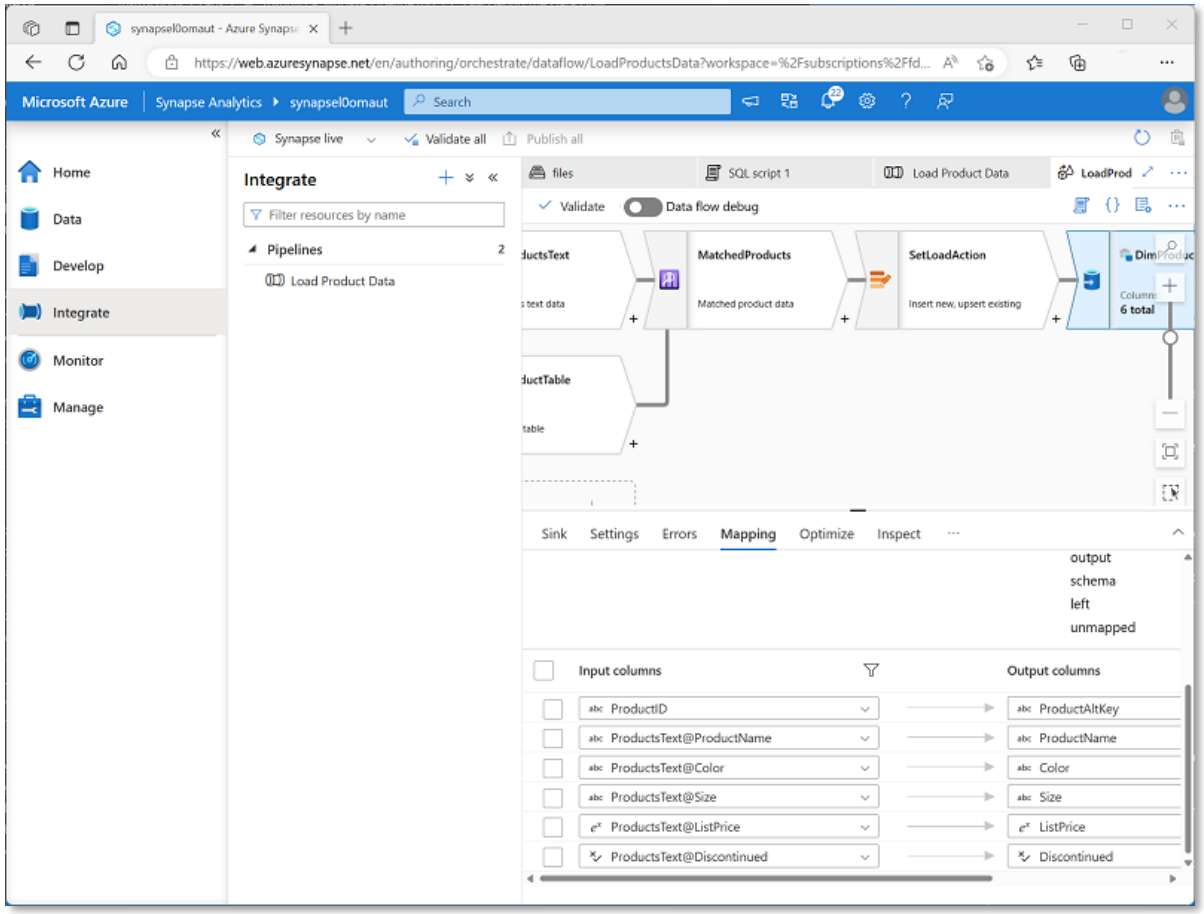
- Create pipelines on the **Integrate** page
- Add and configure activities:
- Specify new or existing datasets and linked services as required in settings

  They'll be added to the **Data** and **Manage** pages
- Connect activities to define processing flow – define paths for:
  - Succeeded
  - Failed
  - Completed



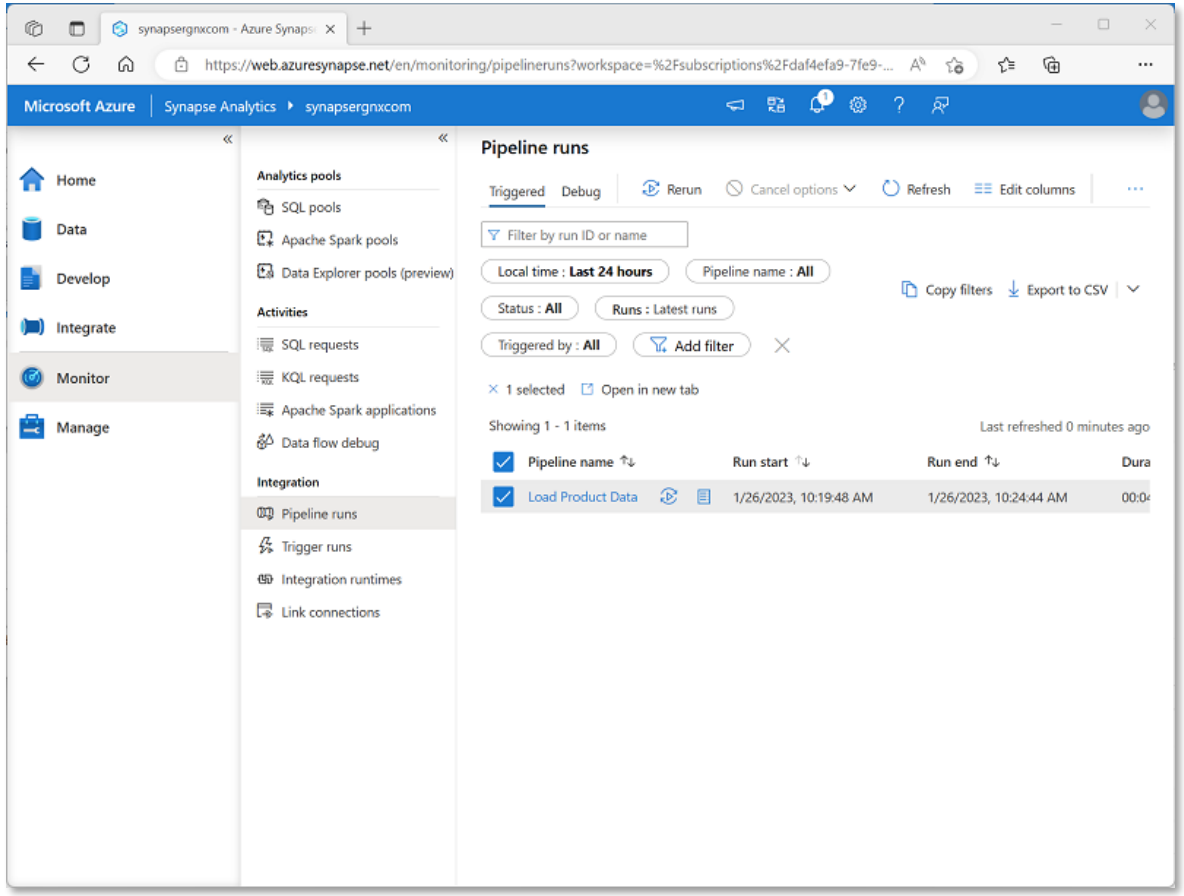
Define data flows

- A **Data Flow** is a commonly used activity type to define data flow and transformation
- Consists of:
  - **Sources** - Data sets that map to data stores
  - **Transformations** – operations on data as it streams through the data flow
  - **Sinks** – targets for data to be loaded



Run a pipeline

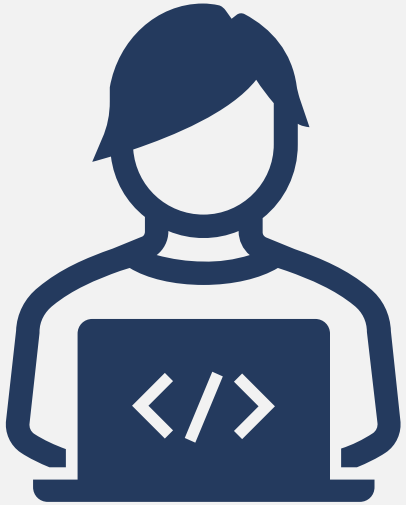
- Debug pipelines to test during development
- Define *triggers* to run pipelines in production:
  - Manual – run immediately
  - Schedule – run at regular intervals
  - Event – run when an event occurs (such as new data saved in a data store)
- Monitor pipeline runs in Azure Synapse Studio



Exercise: Build a data pipeline in Azure Synapse Analytics

Use the hosted lab environment provided, or view the lab instructions at the link below:

https://aka.ms/mslearn-build-synapse-pipeline



Knowledge check

What does a pipeline use to access external data source and processing resources?

- ❑ Data Explorer pools
- ☑ Linked services
- ❑ External tables

What kind of object should you add to a data flow to define a target to which data is loaded?

- ❑ Source
- ❑ Transformation
- ☑ Sink

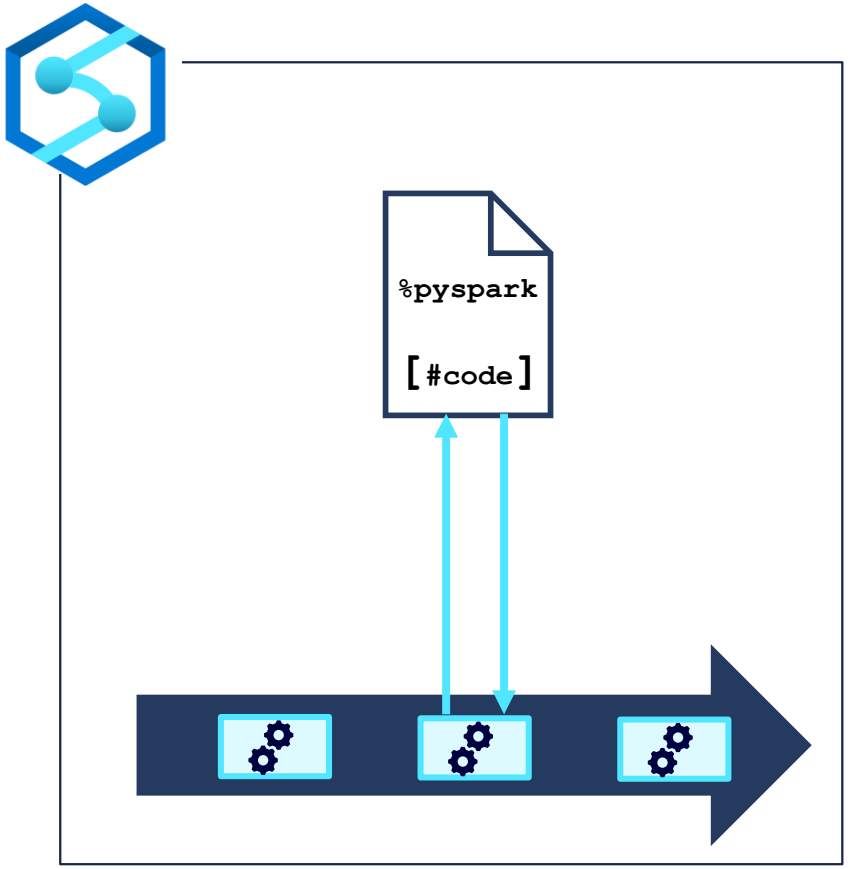
What must you create to run a pipeline at scheduled intervals?

- ❑ A control flow
- ☑ A trigger
- ❑ An activity

Use Spark Notebooks in an Azure Synapse Pipeline

Synapse notebooks and pipelines

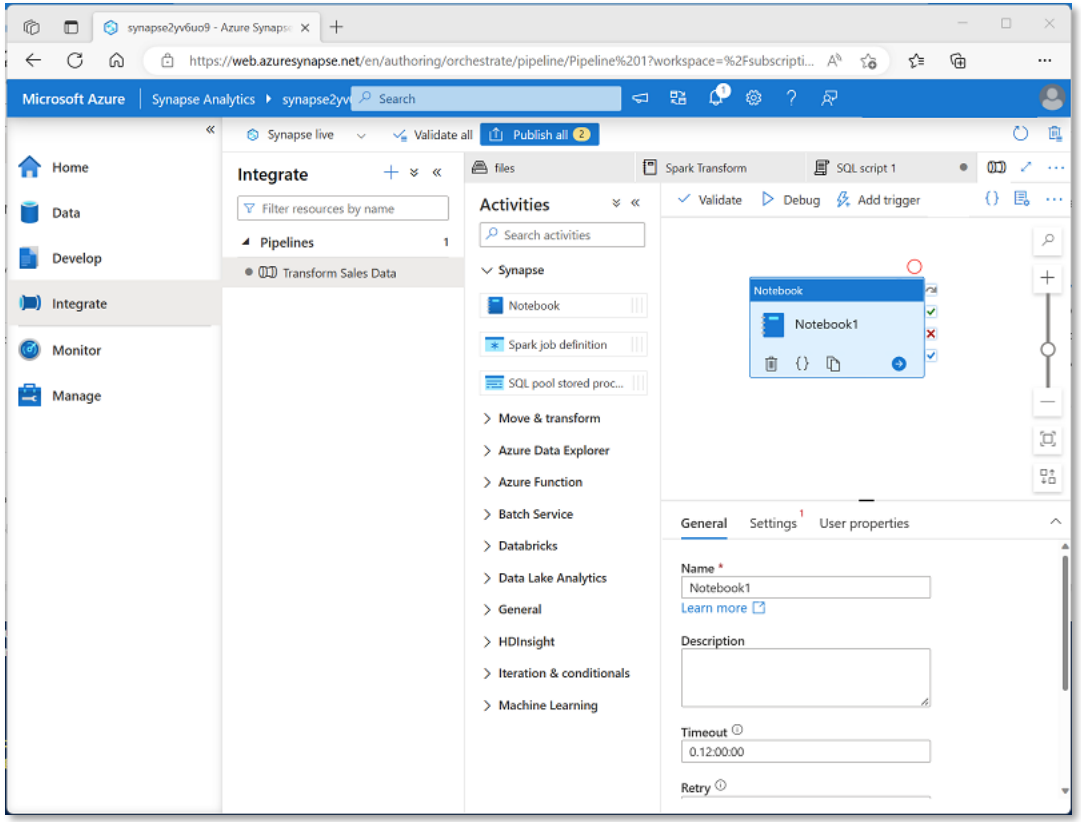
- Use Synapse notebooks to develop and test data transformation code on Apache Spark
- Incorporate notebooks into data ingestion and transformation pipelines
- Notebooks run in the specified Spark pool in the Synapse workspace



Use a Synapse notebook activity in a pipeline

Add a *Notebook* activity to a pipeline, specifying:

- **General** properties such as name, timeout, and number of retries
- **Settings**, such as the notebook to be run, the spark pool on which to run it, and parameter values
- **User properties** to define custom configuration values



Use parameters in a notebook

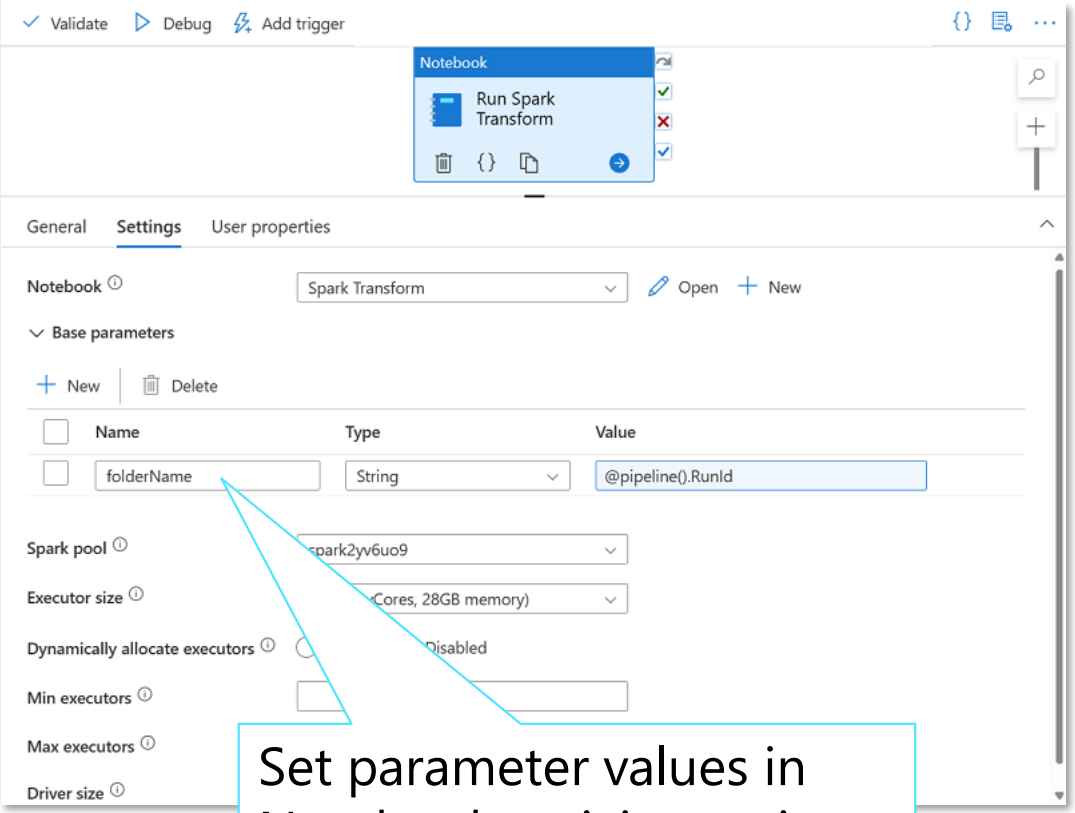
```
import uuid

# Variable for unique folder name
folderName = uuid.uuid4()
```

Press shift + enter to run

Parameters

Declare variables in a *Parameters* cell in the notebook, with a default value

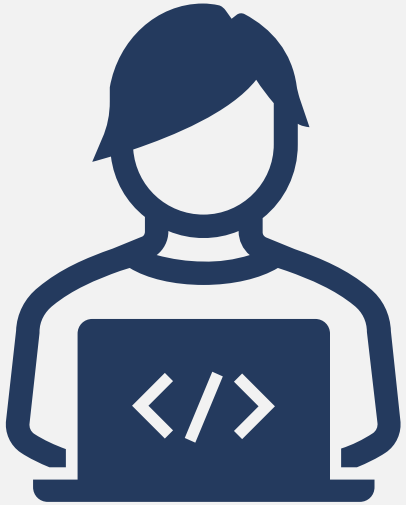


Set parameter values in Notebook activity settings

Exercise: Use an Apache Spark notebook in a pipeline

Use the hosted lab environment provided, or view the lab instructions at the link below:

https://aka.ms/mslearn-spark-synapse-pipeline



Knowledge check

What kind of pool is required to run a Synapse notebook in a pipeline?

- ❑ A *Dedicated SQL* pool
- ❑ A *Data Explorer* pool
- ☑ An *Apache Spark* pool

What kind of pipeline activity encapsulates a Synapse notebook?

- ☑ Notebook activity
- ❑ HDInsight Spark activity
- ❑ Script activity

A notebook cell contains variable declarations. How can you use these as parameters?

- ❑ Add a *%%Spark* magic at the beginning of the cell
- ☑ Toggle the *Parameters cell* setting for the cell
- ❑ Use the *var* keyword for each variable declaration

Further reading

Transfer and transform data with Azure Synapse Analytics Pipelines
https://aka.ms/mslearn-synapse-pipelines