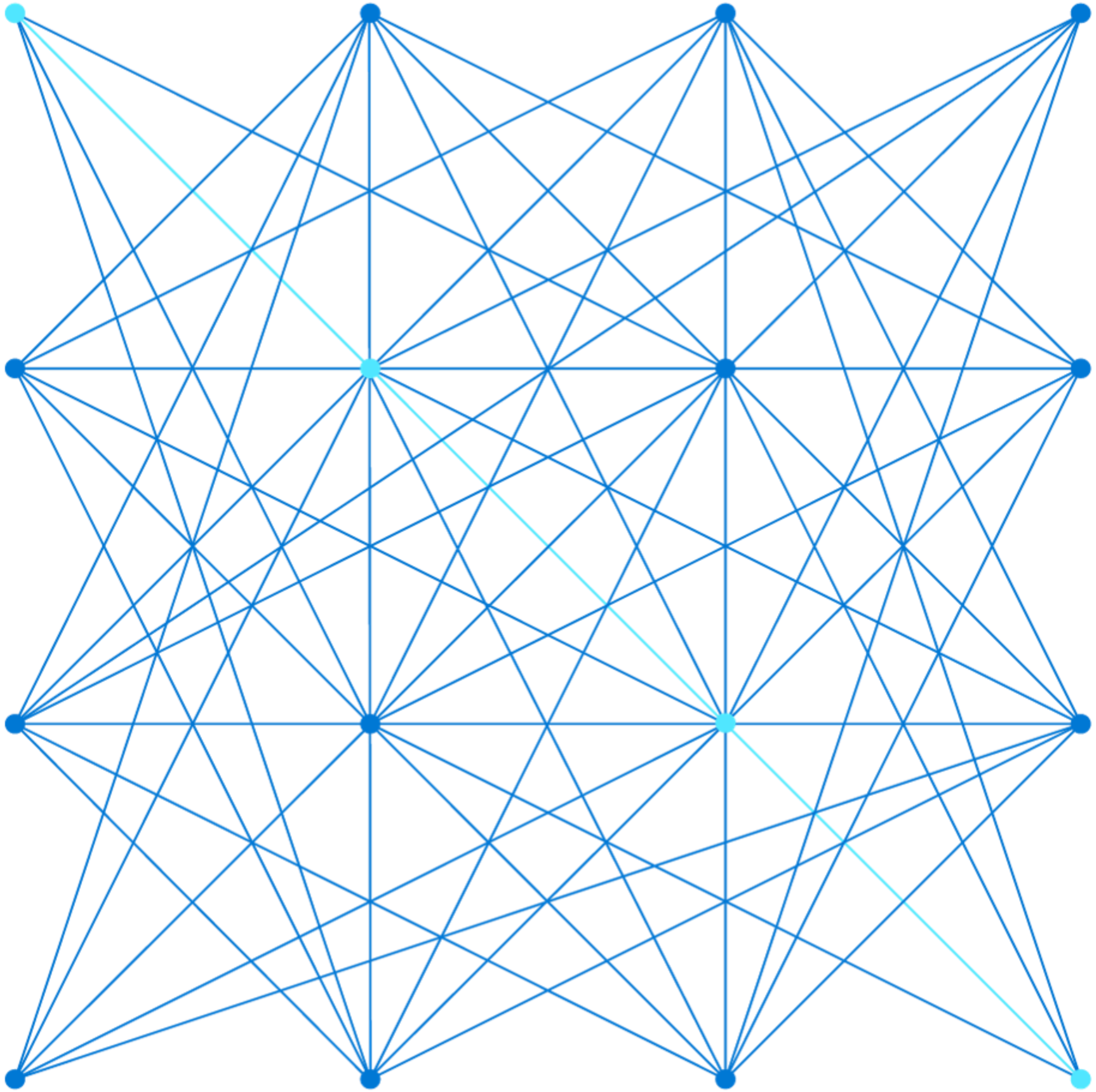
Microsoft Azure

DP-203T00: Explore, transform, and load data into the Data Warehouse using Apache Spark



Agenda

Lesson 01 – Understand big data engineering with Apache Spark in Azure Synapse Analytics

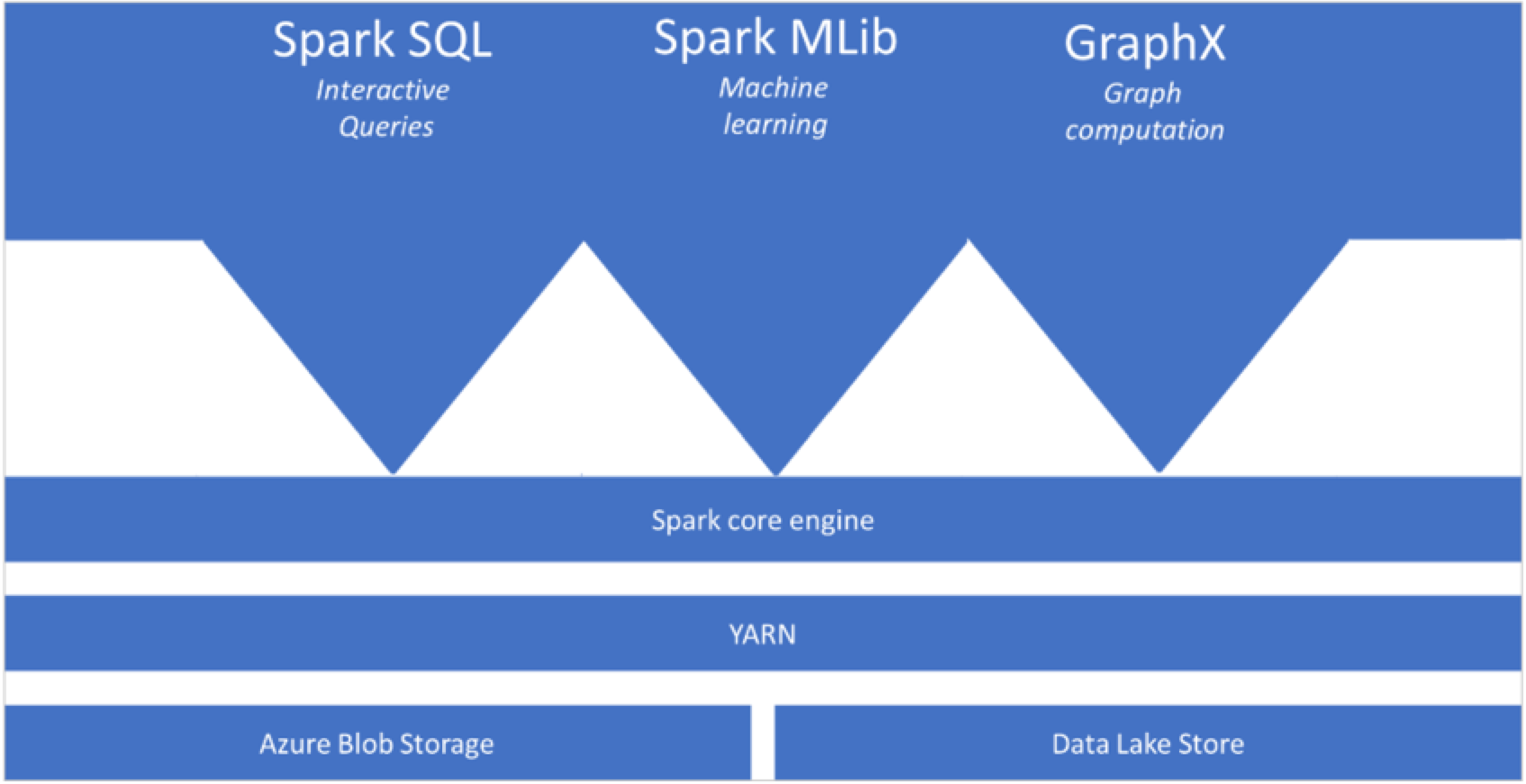
Lesson 02 – Ingest data with Apache Spark notebooks in Azure Synapse Analytics

Lesson 03 – Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics

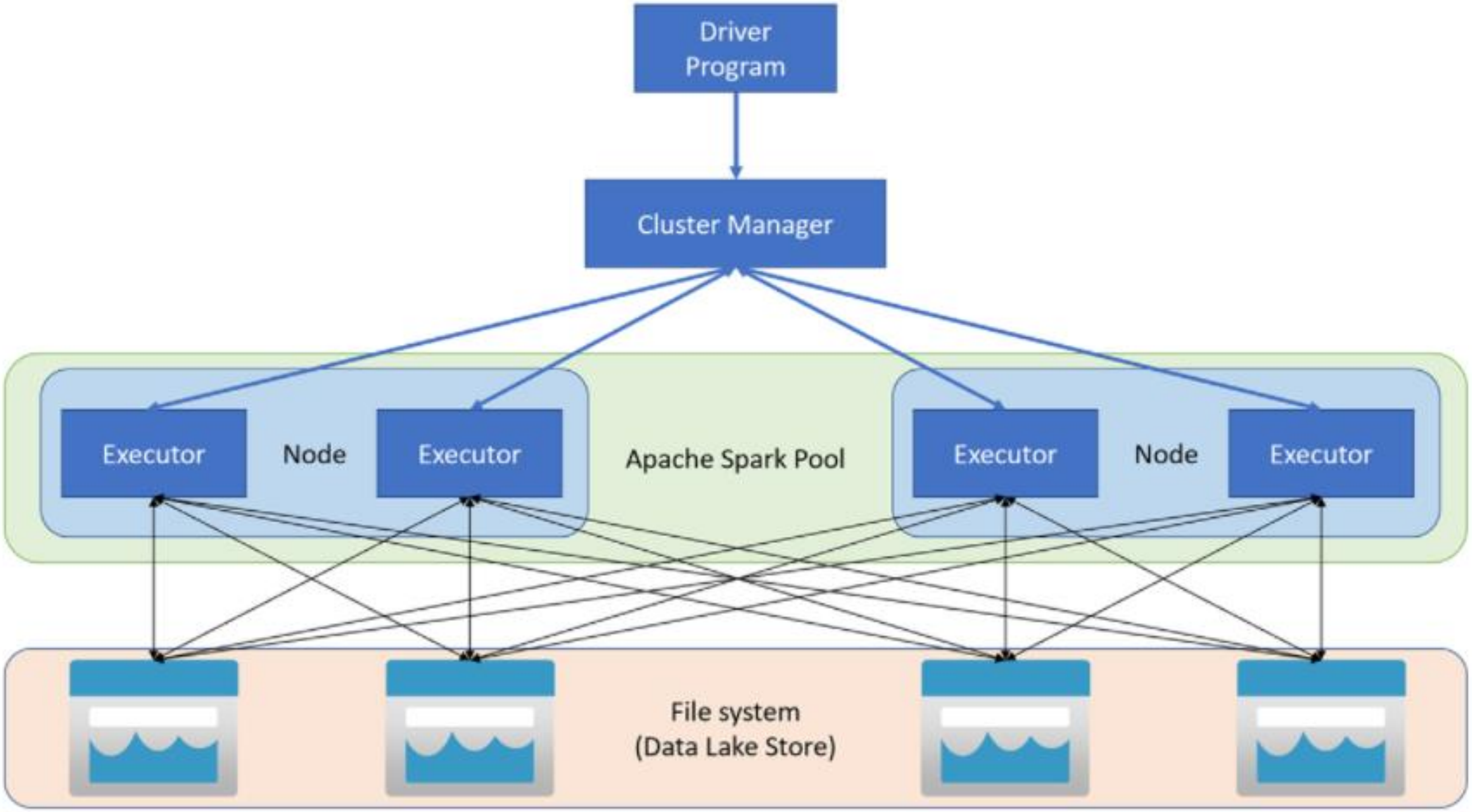
Lesson 04 – Integrate SQL and Apache Spark pools in Azure Synapse Analytics

Lesson 01: Understand big data engineering with Apache Spark in Azure Synapse Analytics

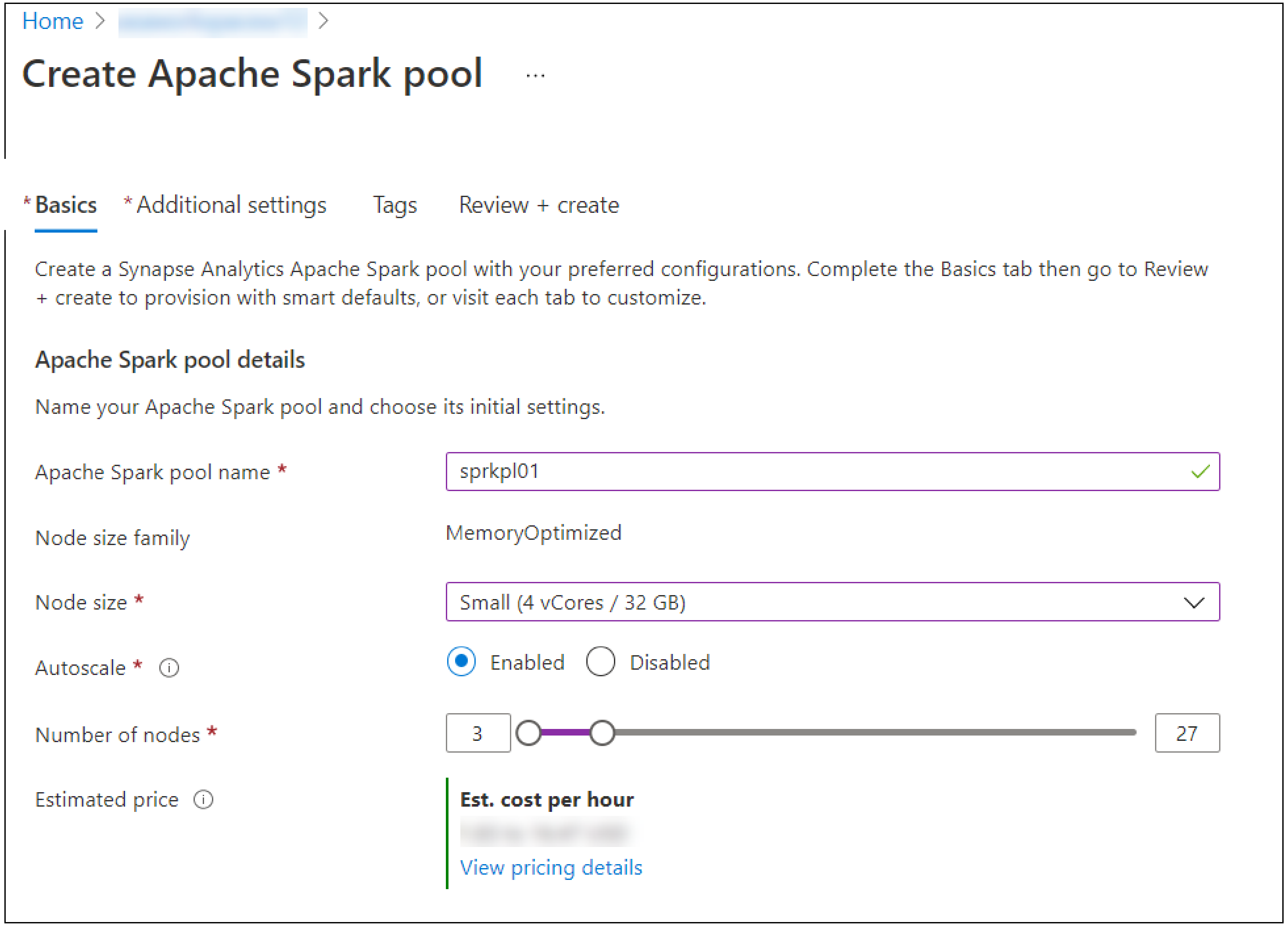
Introduction to big data engineering with Apache Spark in Azure Synapse Analytics



How do Apache Spark pools work in Azure Synapse Analytics



How to create an Apache Spark pool in Azure Synapse Analytics



Home > [redacted] >

Create Apache Spark pool ...

Create a Synapse Analytics Apache Spark pool with your preferred configurations. Complete the Basics tab then go to Review + create to provision with smart defaults, or visit each tab to customize.

Apache Spark pool details

Name your Apache Spark pool and choose its initial settings.

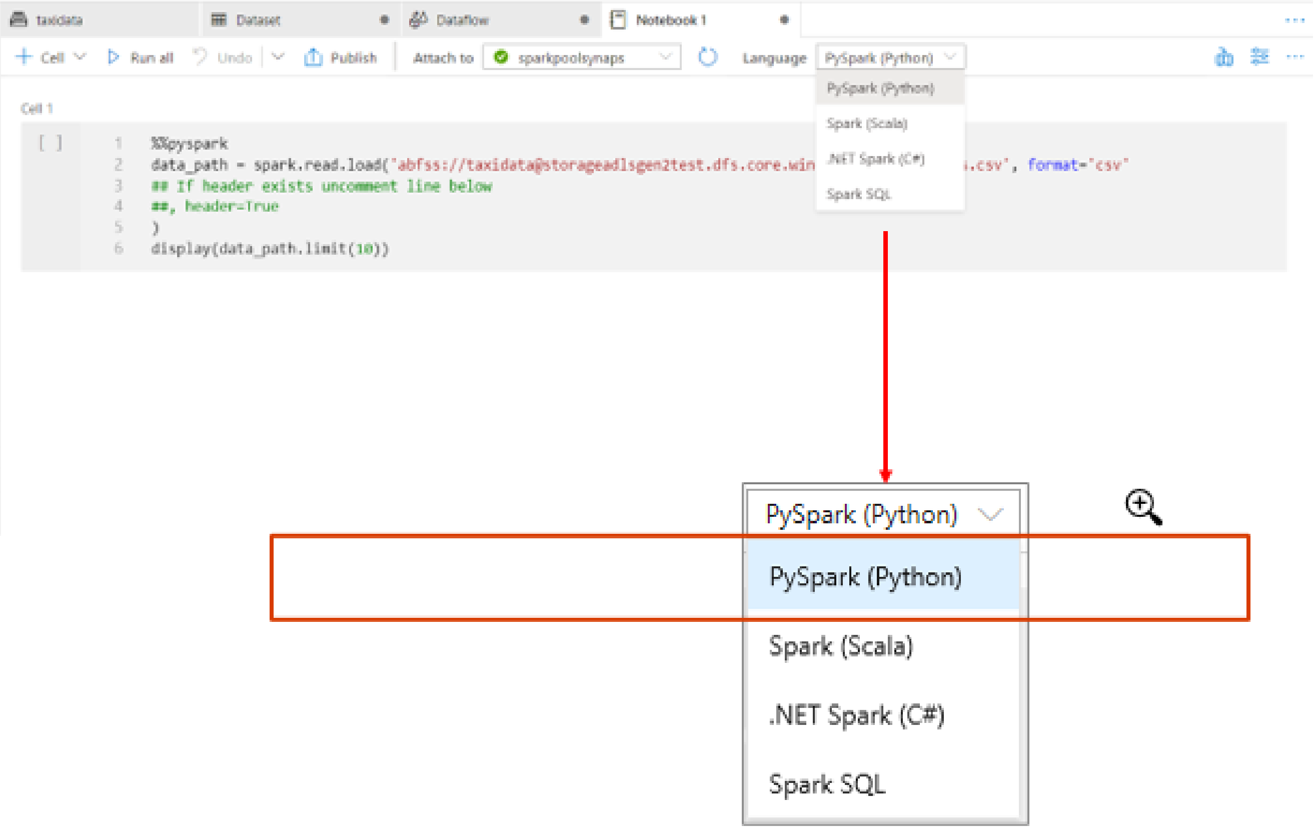
Apache Spark pool name *	sprkpl01 ✓
Node size family	MemoryOptimized
Node size *	Small (4 vCores / 32 GB)
Autoscale * ⓘ	● Enabled ○ Disabled
Number of nodes *	3 – 27
Estimated price ⓘ	**Est. cost per hour** [redacted] View pricing details

Lesson 02: Ingest data with Apache Spark notebooks in Azure Synapse Analytics

Apache Spark notebooks features in Azure Synapse Analytics

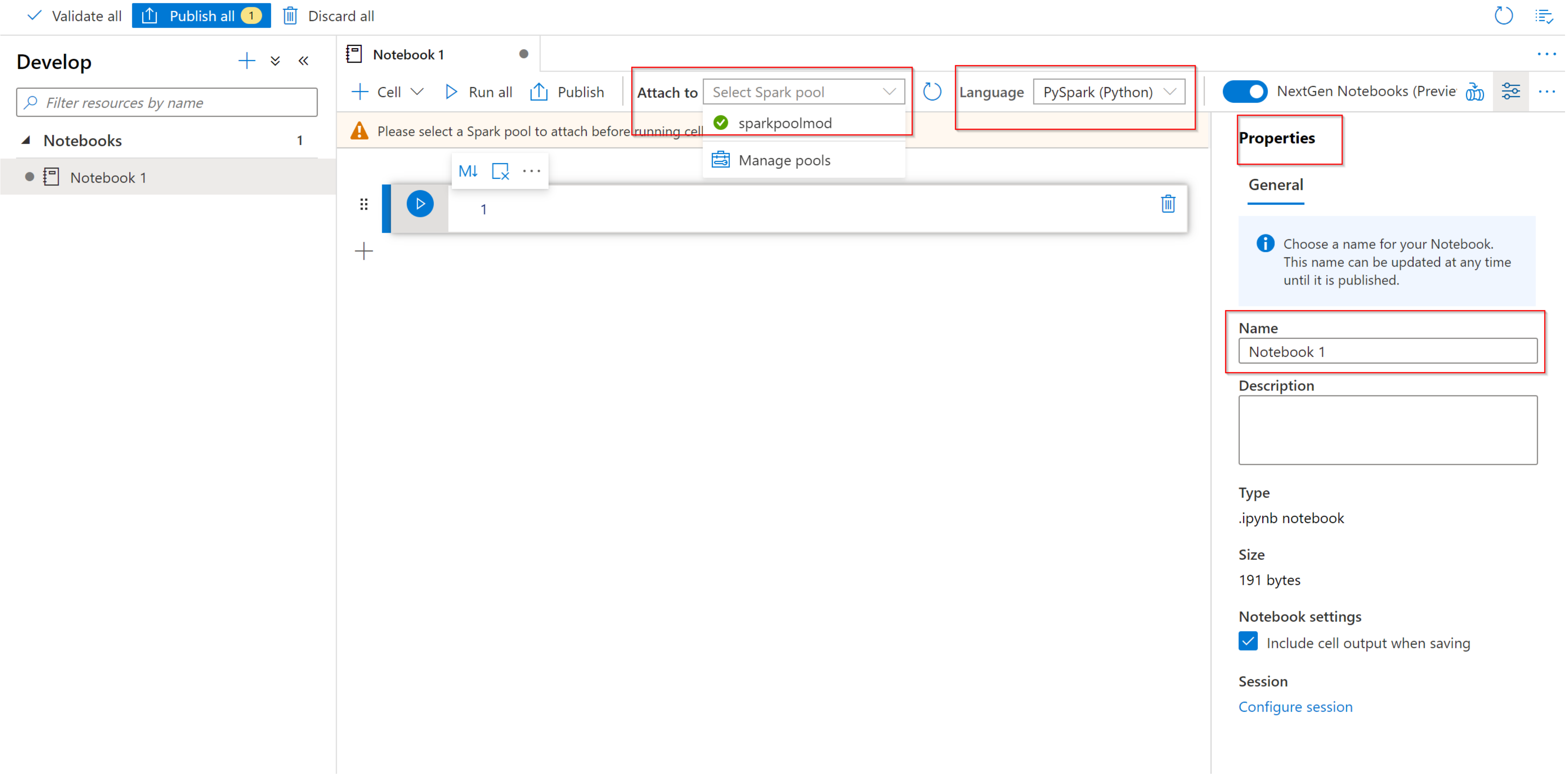
Notebooks

- Access through Synapse Studio
- Examples Available through Knowledge Center
- Allows to write multiple languages in one notebook by using %%<Name of language>
- Support for Language Syntax highlight, syntax error, syntax code completion
- Offers temporary tables across languages
- Export results



```
%%pyspark
data_path = spark.read.load('abfss://taxidata@storageadlsgen2test.dfs.core.win          .csv', format='csv'
## If header exists uncomment line below
##, header=True
)
display(data_path.limit(10))
```

Creating a notebook in Azure Synapse Analytics



Attach to: Select Spark pool

sparkpoolmod

Language: PySpark (Python)

Please select a Spark pool to attach before running cell

Choose a name for your Notebook. This name can be updated at any time until it is published.

Name: Notebook 1

Description:

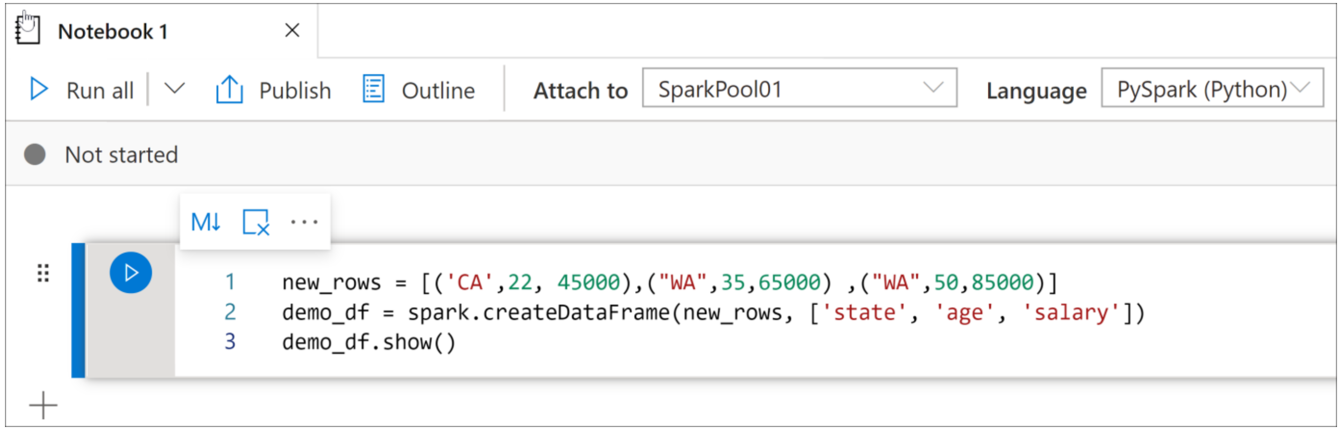
Type: .ipynb notebook

Size: 191 bytes

☑ Include cell output when saving

Ingest data with Apache Spark notebooks in Azure Synapse Analytics

- Generating data while executing the command

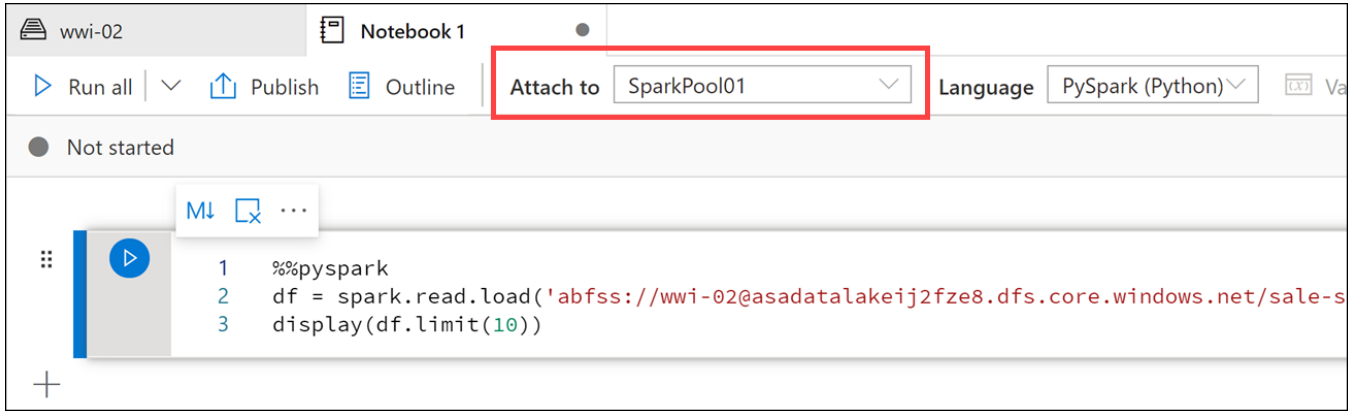


Attach to SparkPool01

Language PySpark (Python)

```
new_rows = [('CA',22, 45000),("WA",35,65000) ,("WA",50,85000)]
demo_df = spark.createDataFrame(new_rows, ['state', 'age', 'salary'])
demo_df.show()
```

- Loading data in a single command from a data file



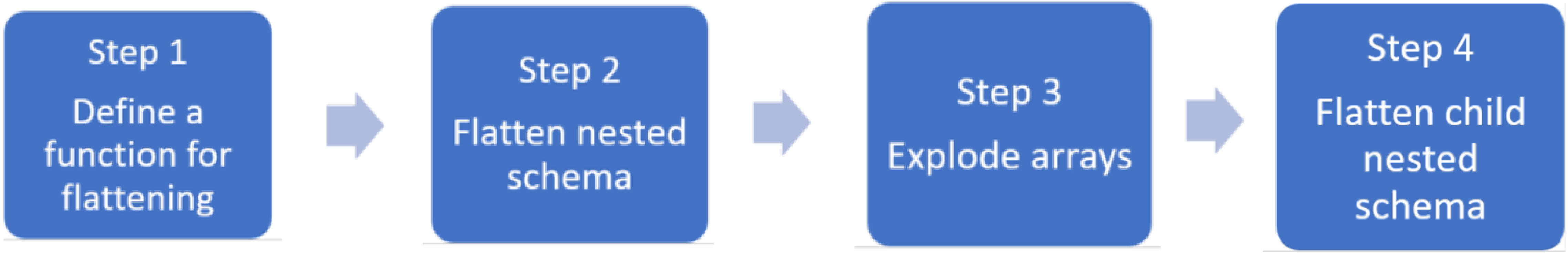
Attach to SparkPool01

Language PySpark (Python)

```
%%pyspark
df = spark.read.load('abfss://wwi-02@asadatalakeij2fze8.dfs.core.windows.net/sale-s
display(df.limit(10))
```

Lesson 03: Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics

Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics



Step 1

Define a function for flattening

Step 2

Flatten nested schema

Step 3

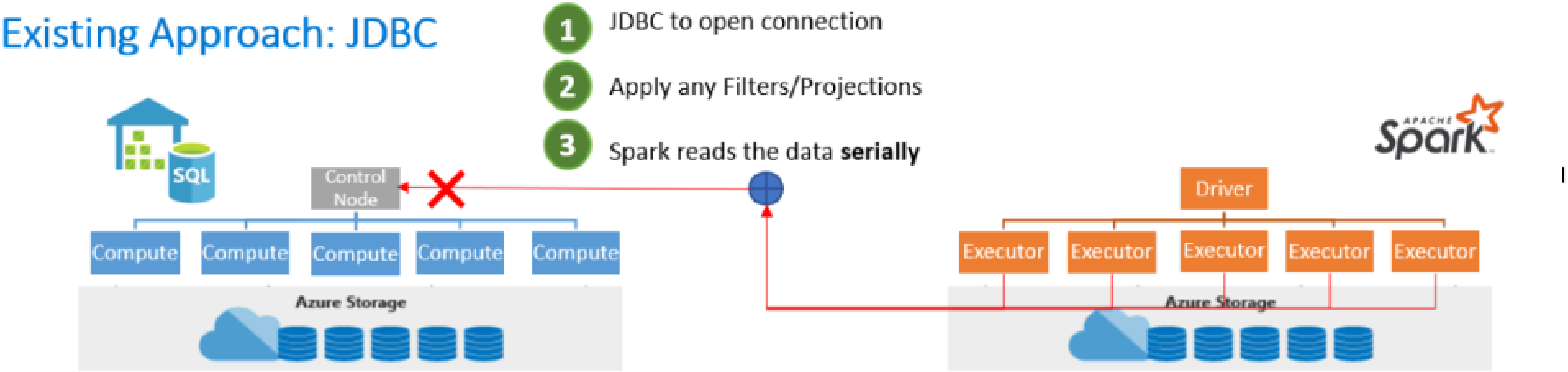
Explode arrays

Step 4

Flatten child nested schema

Lesson 04: Integrate SQL and Apache Spark pools in Azure Synapse Analytics

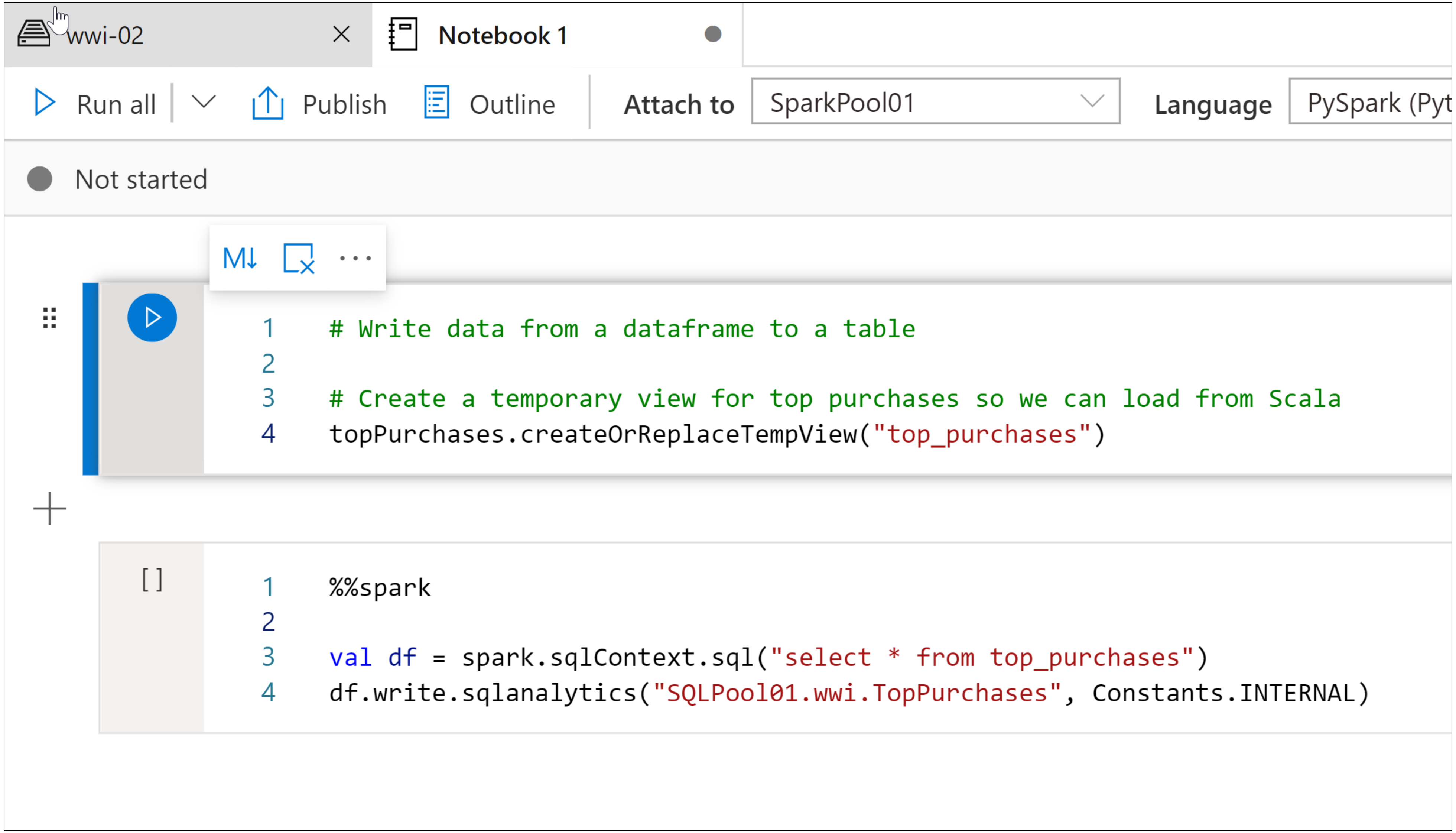
Integrate SQL and Apache Spark pools in Azure Synapse Analytics



Existing Approach: JDBC

1. JDBC to open connection
2. Apply any Filters/Projections
3. Spark reads the data **serially**

Write data from Apache Spark pools to a dedicated SQL pool



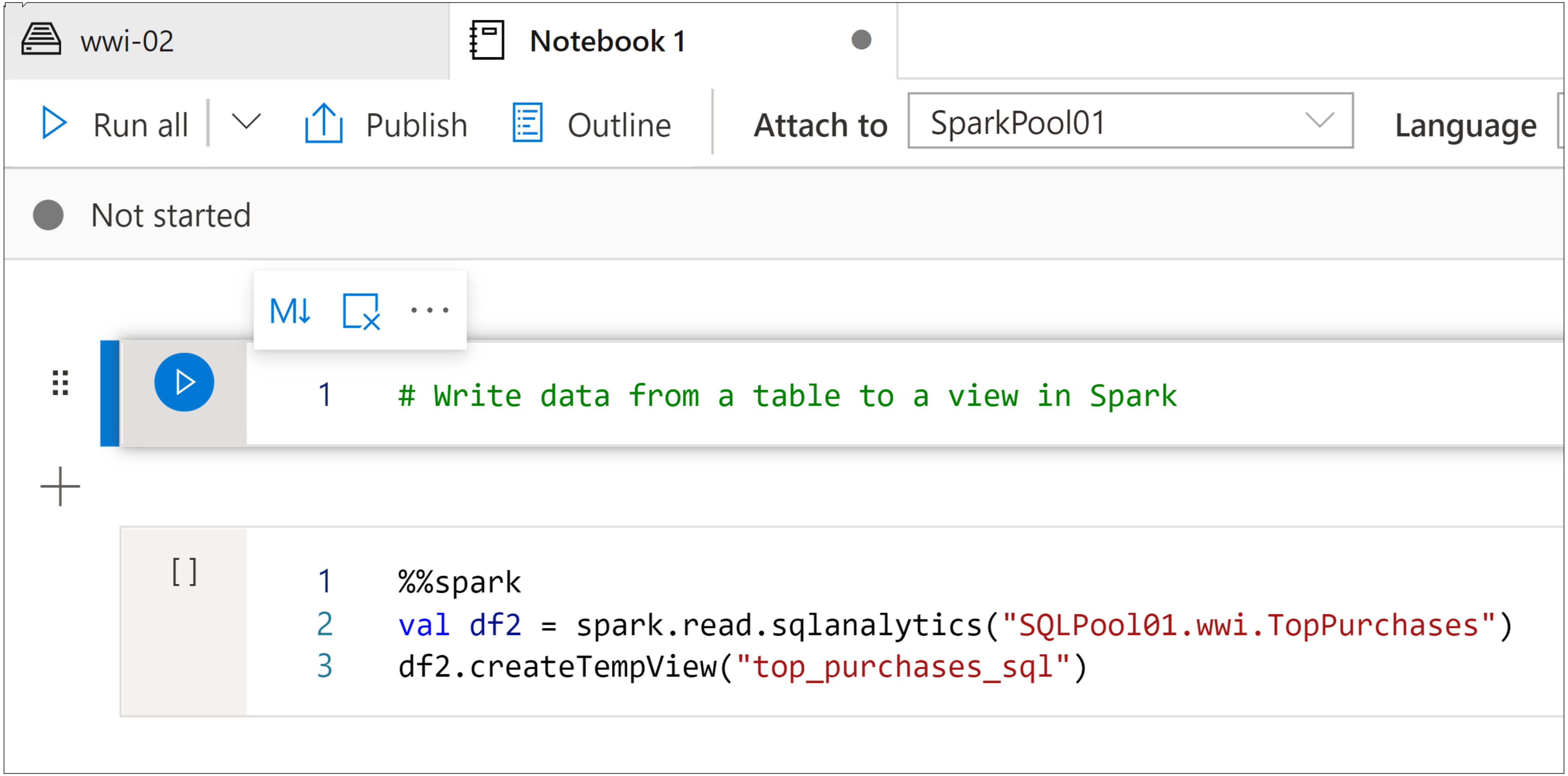
```
# Write data from a dataframe to a table

# Create a temporary view for top purchases so we can load from Scala
topPurchases.createOrReplaceTempView("top_purchases")
```

```
%%spark

val df = spark.sqlContext.sql("select * from top_purchases")
df.write.sqlanalytics("SQLPool01.wwi.TopPurchases", Constants.INTERNAL)
```

Write data from a dedicated SQL pool to Apache Spark pools



```
# Write data from a table to a view in Spark
```

```
%%spark
val df2 = spark.read.sqlanalytics("SQLPool01.wwi.TopPurchases")
df2.createTempView("top_purchases_sql")
```

Review questions

Q01 – What is an element of a Spark Pool in Azure Synapse Analytics?

A01 – Spark Instance

Q02 – How can all Apache Spark notebooks in Synapse Studio be saved?

A02 – Select the Publish all button on the workspace command bar.

Q03 – When is it unnecessary to use import statements for transferring data between a dedicated SQL and Spark pool?

A03 – Use the integrated notebook experience from Azure Synapse Studio.

Lab: Explore, transform, and load data into the Data Warehouse using Apache Spark

http://fmdk.io/dp2034

Lab overview

This lab teaches you how to explore data stored in a data lake, transform the data, and load data into a relational data store. You will explore Parquet and JSON files and use techniques to query and transform JSON files with hierarchical structures. Then you will use Apache Spark to load data into the data warehouse and join Parquet data in the data lake with data in the dedicated SQL pool.

Lab objectives

After completing this lab, you will be able to:

- Perform Data Exploration in Synapse Studio
- Ingest data with Spark notebooks in Azure Synapse Analytics
- Transform data with DataFrames in Spark pools in Azure Synapse Analytics
- Integrate SQL and Spark pools in Azure Synapse Analytics

Lab review

Q01 – Which command is used to analyze parquet files and infer schema's using the Spark Engine?

Q02 – What is an option to you query JSON files using the SQL syntax in an Apache Spark Notebook connected to a Spark Pool in Azure Synapse Analytics?

Q03 – How do you set the language of a cell in an Apache Spark Notebook?

Module summary

In this module, you have learned about:

- Azure Synapse Analytics
- Apache Spark Notebooks
- Integration of SQL and Spark
- DataFrames
- Apache Spark Architecture

Next steps

After the course, consider visiting [Azure Apache Spark for Azure Synapse Analytics]. The Apache Spark in Azure Synapse Analytics provides an overview of how Apache Spark is integrated with Azure Synapse Analytics.

Microsoft Azure