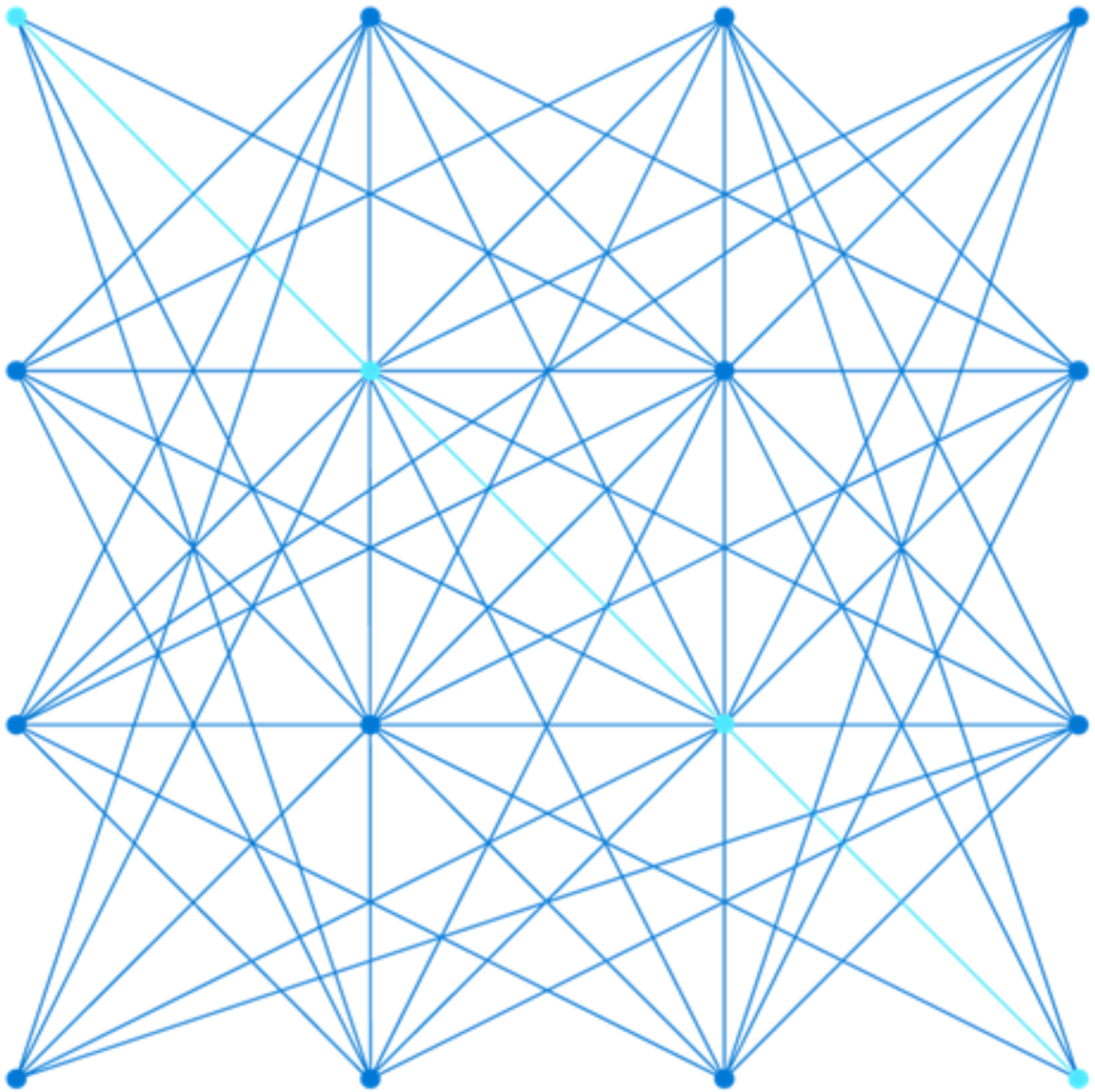
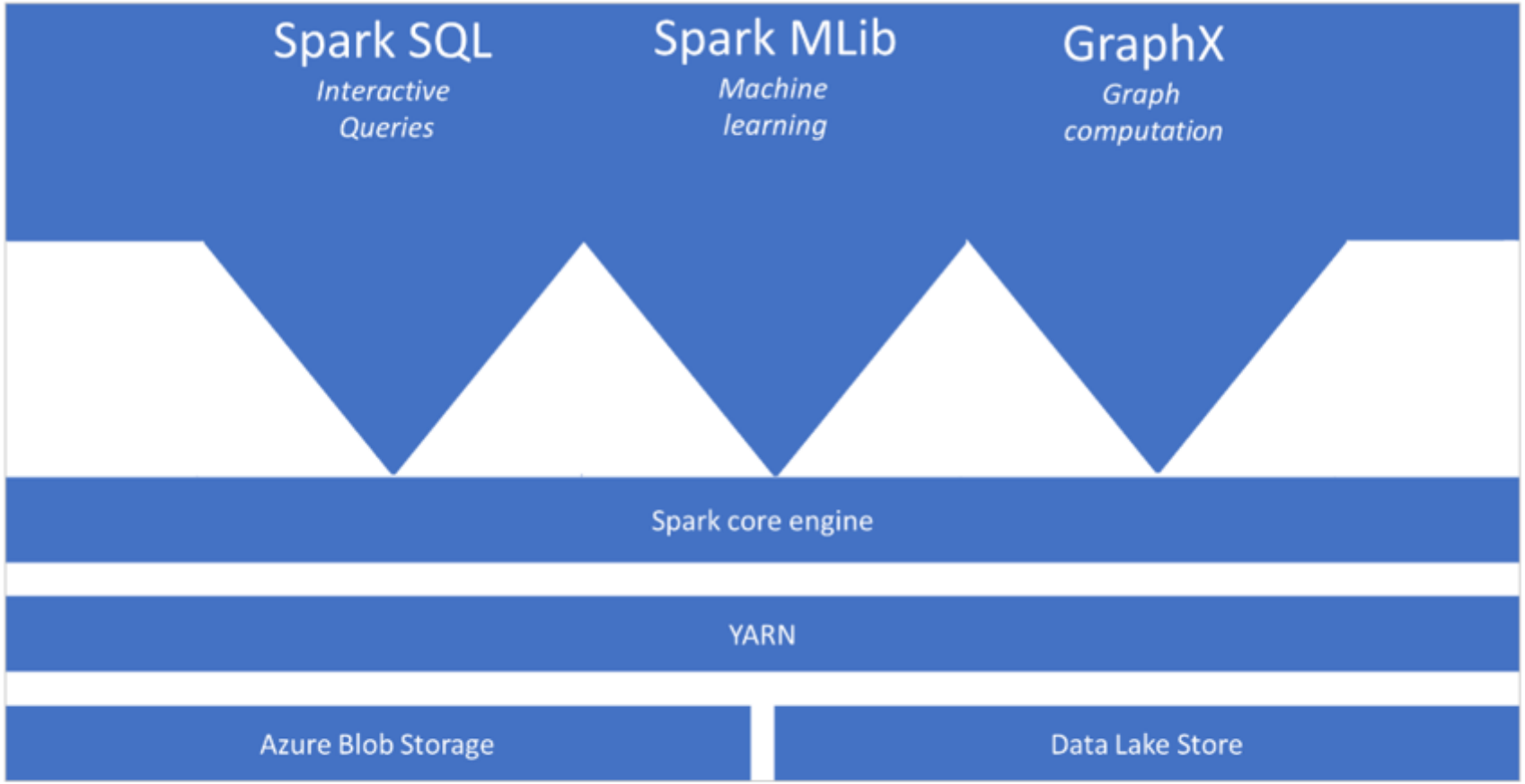
Microsoft Azure

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

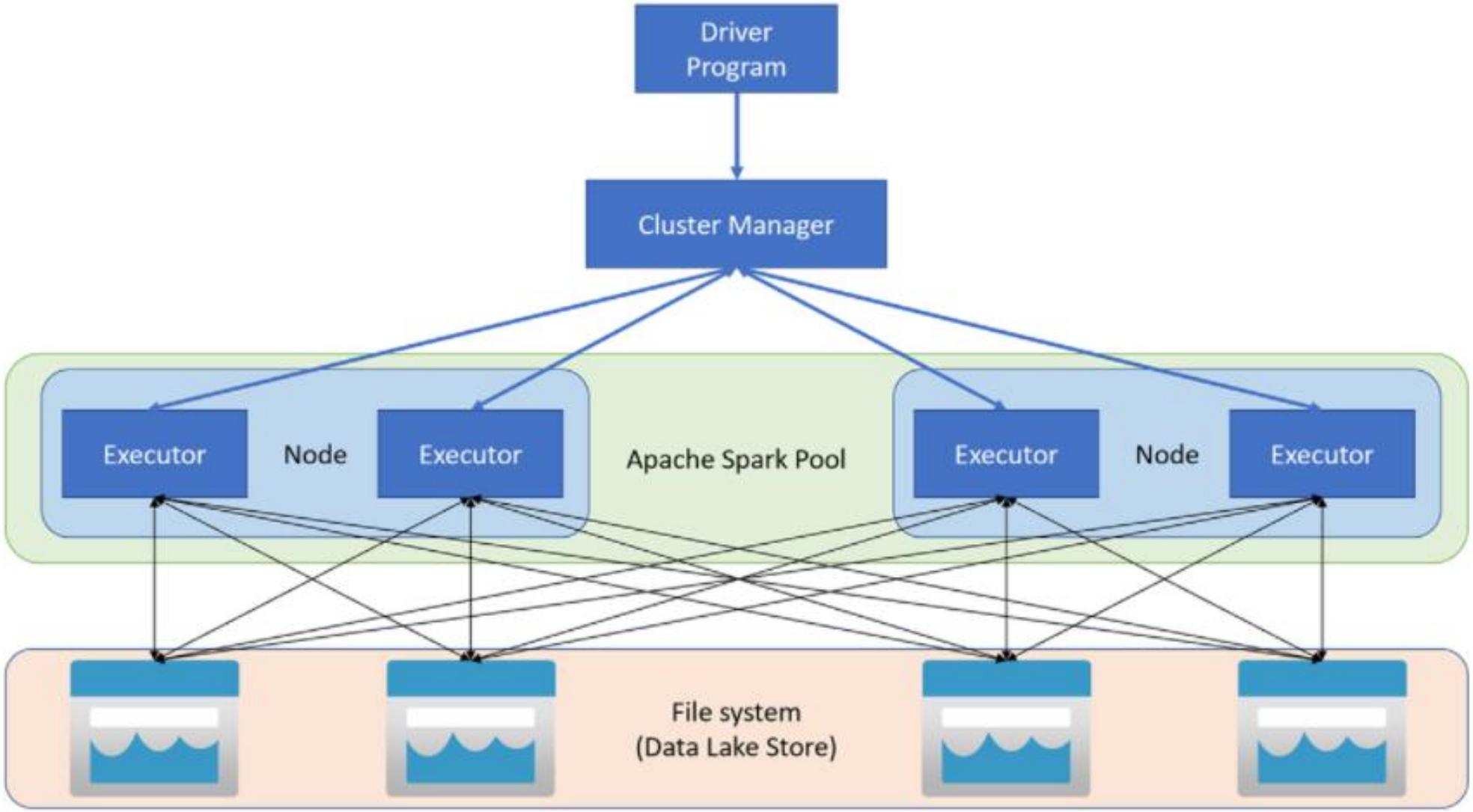


Lesson 01: Understand big data engineering with Apache Spark in Azure Synapse Analytics

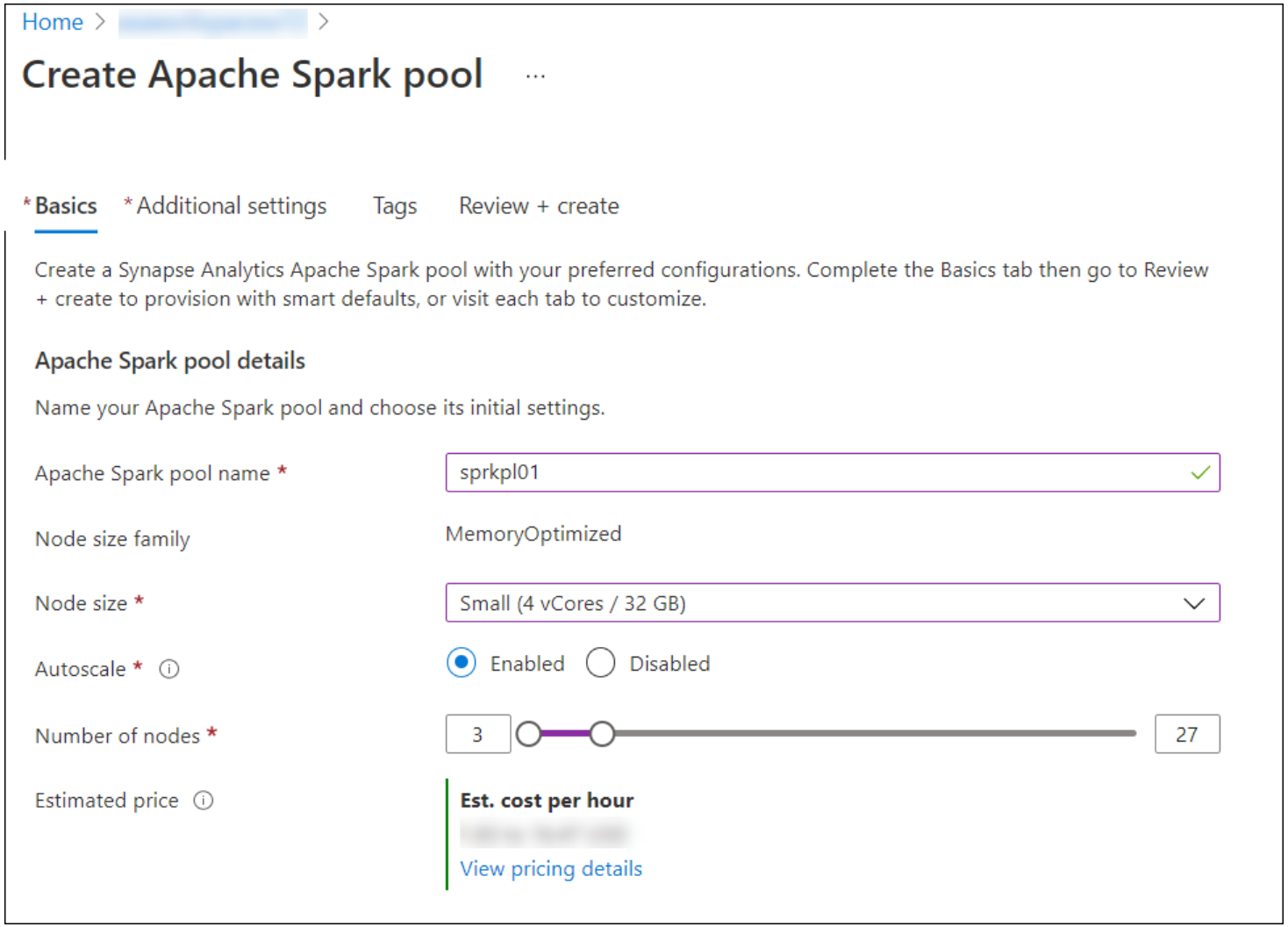
Introduction to big data engineering with Apache Spark in Azure Synapse Analytics



How do Apache Spark pools work in Azure Synapse Analytics



How to create an Apache Spark pool in Azure Synapse Analytics

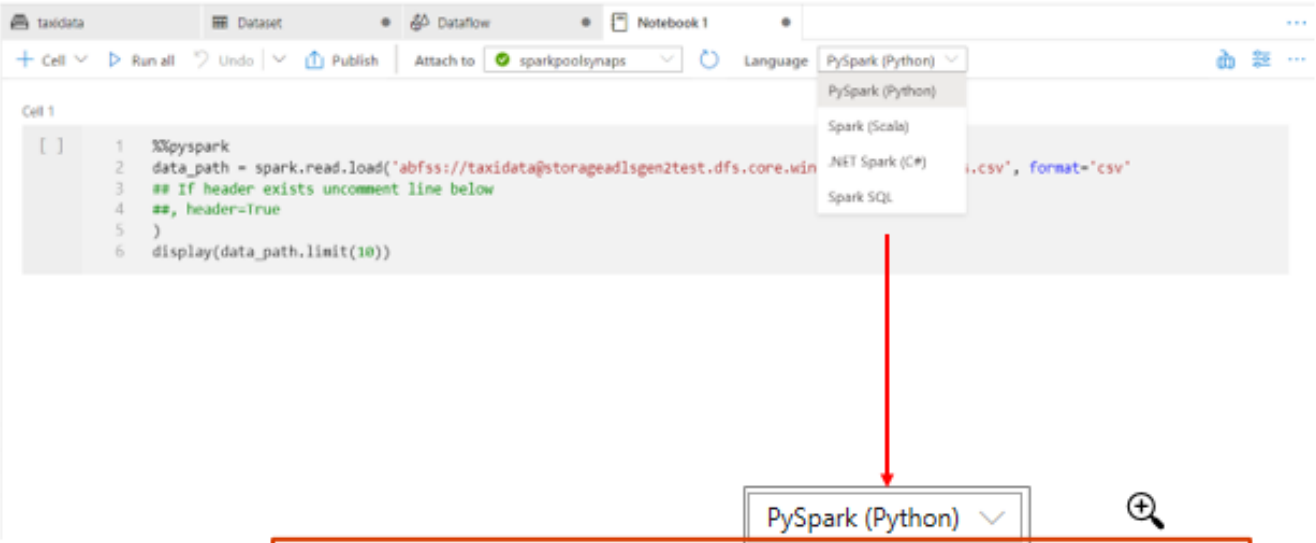


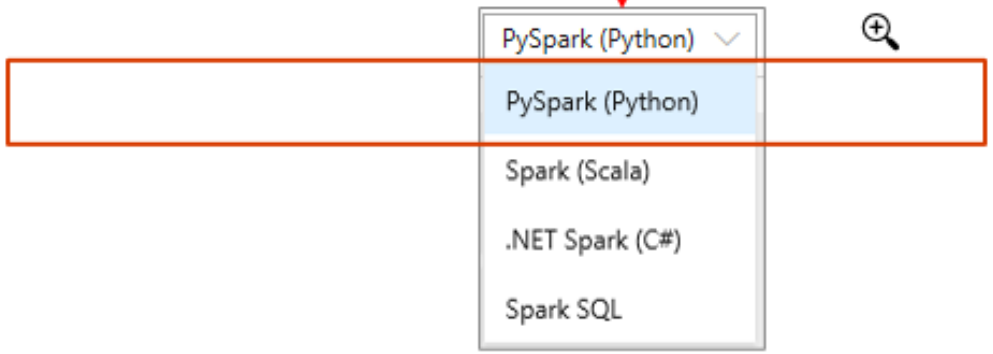
Lesson 02: Ingest data with Apache Spark notebooks in Azure Synapse Analytics

Apache Spark notebooks features in Azure Synapse Analytics

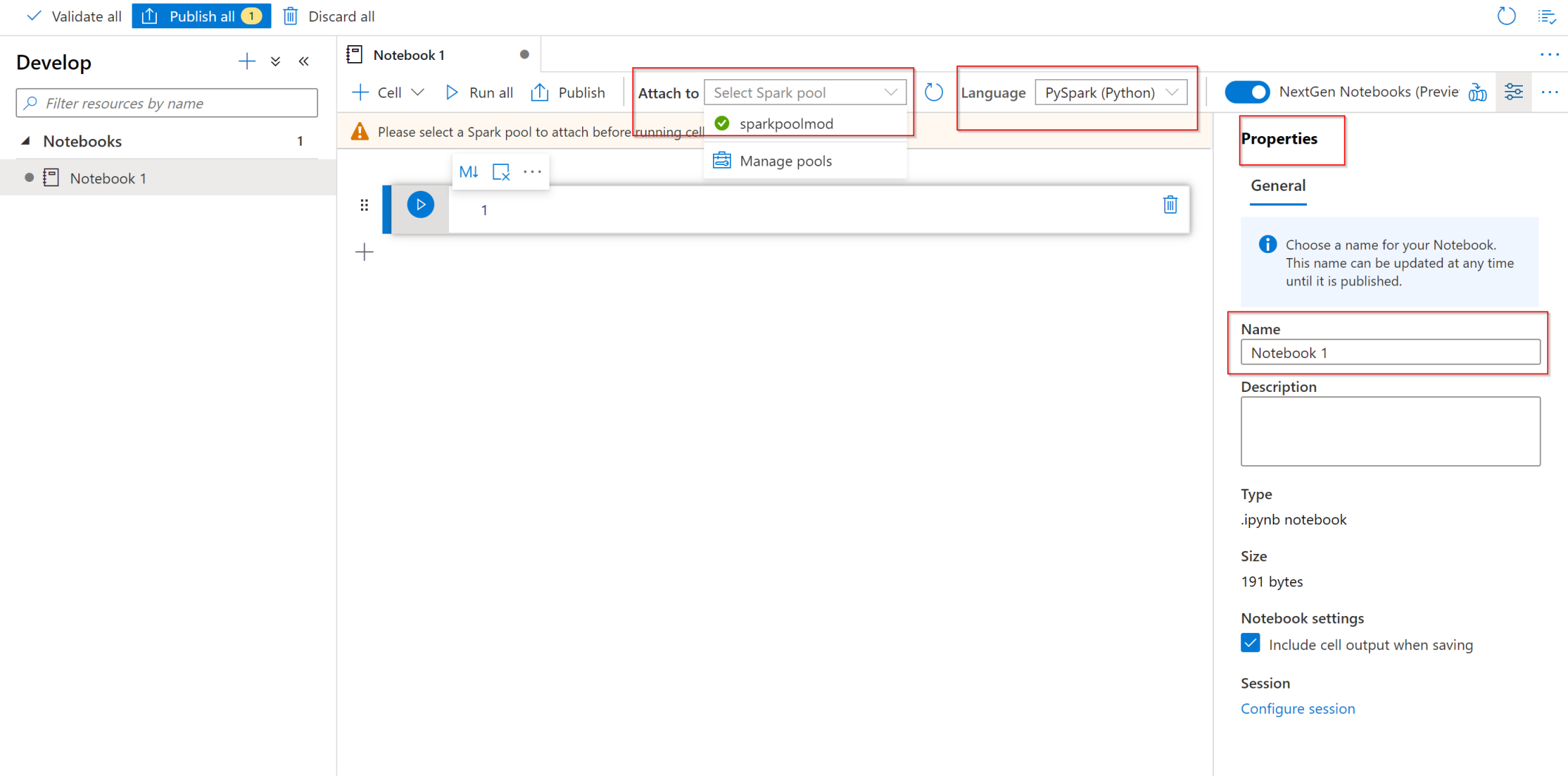
Notebooks

- 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



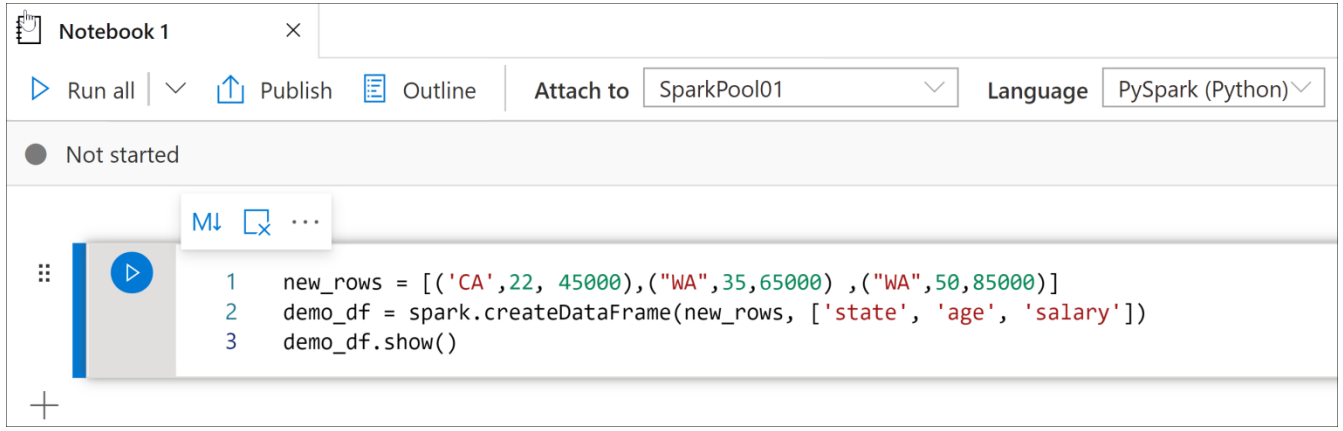


Creating a notebook in Azure Synapse Analytics

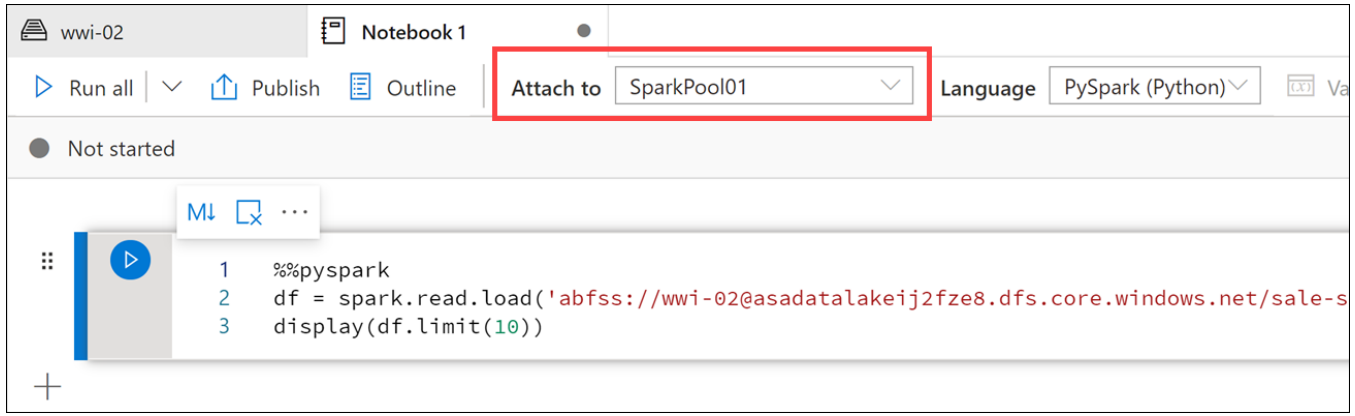


Ingest data with Apache Spark notebooks in Azure Synapse Analytics

- Generating data while executing the command

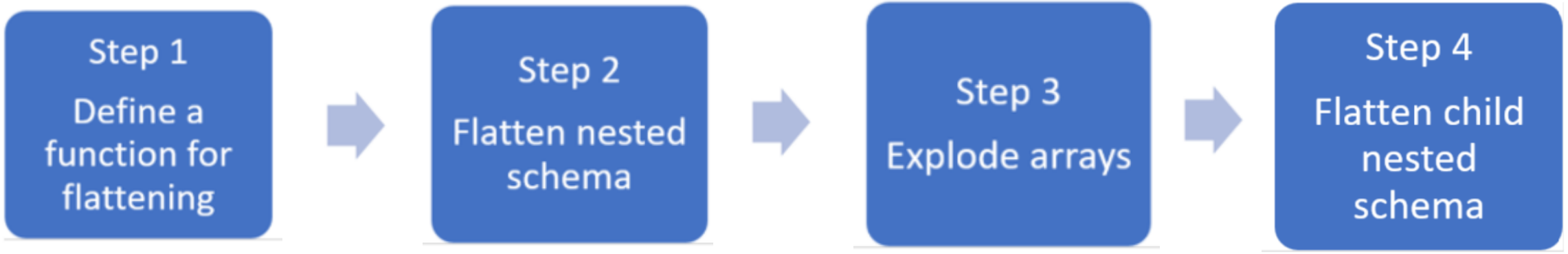


- Loading data in a single command from a data file



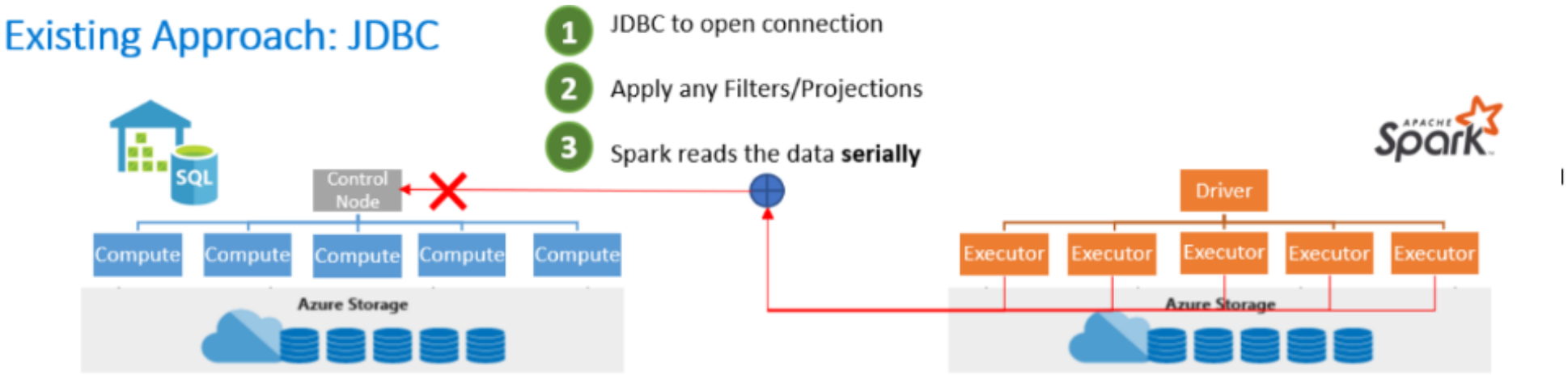
Lesson 03: Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics

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



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

Integrate SQL and Apache Spark pools in Azure Synapse Analytics



Write data from Apache Spark pools to a dedicated SQL pool

wwi-02 | Notebook 1

Run all | Publish | Outline | Attach to SparkPool01 | Language PySpark (Pyt

Not started

```
# 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

wwi-02 | Notebook 1

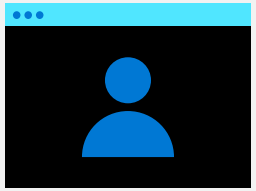
Run all | Publish | Outline | Attach to SparkPool01 | Language

Not started

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

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

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



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