# Transform data using Spark in Synapse Analytics

Data engineers often use Spark notebooks as one of their preferred tools to perform extract, transform, and load (ETL) or extract, load, and transform (ELT) activities that transform data from one format or structure to another.

In this exercise, you’ll use a Spark notebook in Azure Synapse Analytics to transform data in files.

This exercise should take approximately ****30**** minutes to complete.

## Provision an Azure Synapse Analytics workspace

1. In the PowerShell pane, enter the following commands to clone this repo:

rm -r dp-203 -f

git clone https://github.com/MicrosoftLearning/dp-203-azure-data-engineer dp-203

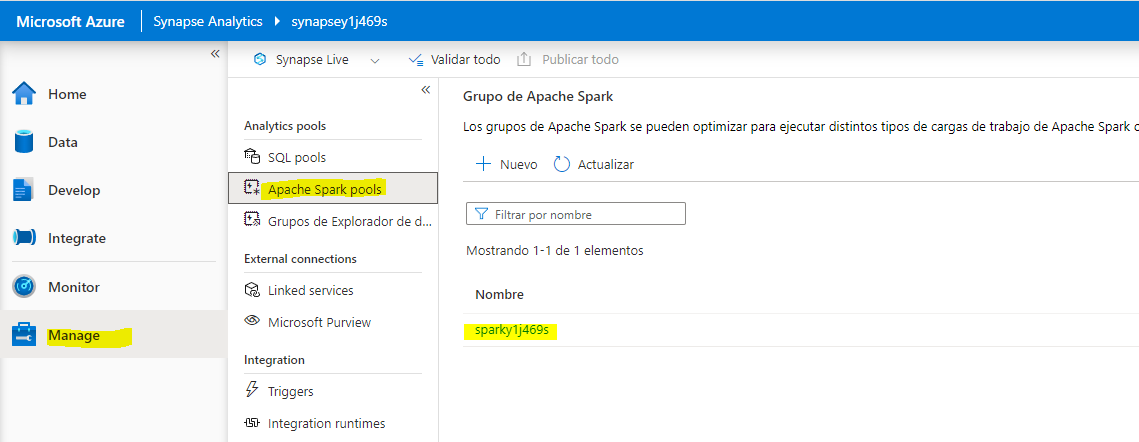
1. After the repo has been cloned, enter the following commands to change to the folder for this exercise and run the ****setup.ps1**** script it contains:

cd dp-203/Allfiles/labs/06

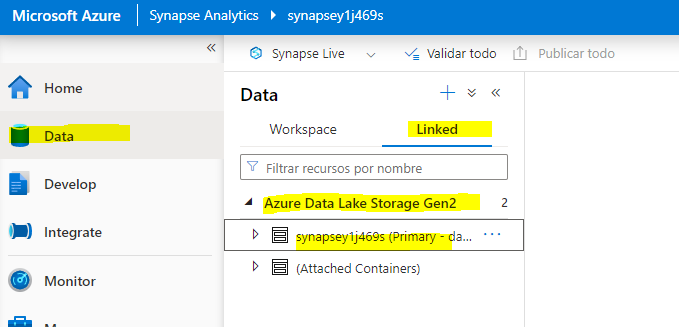
./setup.ps1

## Use a Spark notebook to transform data

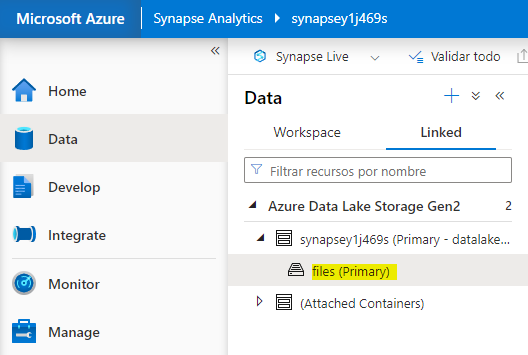
1. After the deployment script has completed, in the Azure portal, go to the ****dp203-**xxxxxxx** resource group that it created, and notice that this resource group contains your Synapse workspace, a Storage account for your data lake, and an Apache Spark pool.
2. Select your Synapse workspace, and in its ****Overview**** page, in the ****Open Synapse Studio**** card, select ****Open**** to open Synapse Studio in a new browser tab; signing in if prompted.
3. On the left side of Synapse Studio, use the ****››**** icon to expand the menu - this reveals the different pages within Synapse Studio that you’ll use to manage resources and perform data analytics tasks.
4. On the ****Manage**** page, select the ****Apache Spark pools**** tab and note that a Spark pool with a name similar to ****spark**xxxxxxx** has been provisioned in the workspace.



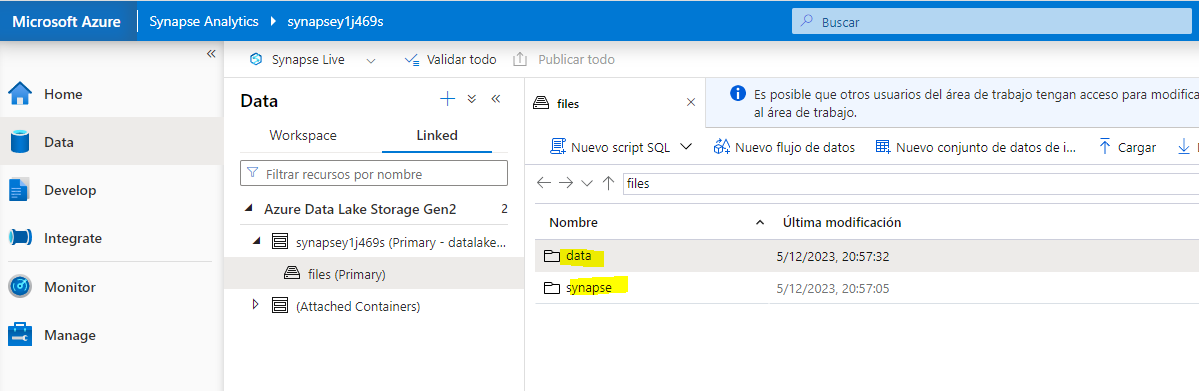
1. On the ****Data**** page, view the ****Linked**** tab and verify that your workspace includes a link to your Azure Data Lake Storage Gen2 storage account, which should have a name similar to ****synapse**xxxxxxx**(Primary - datalake**xxxxxxx**)****.



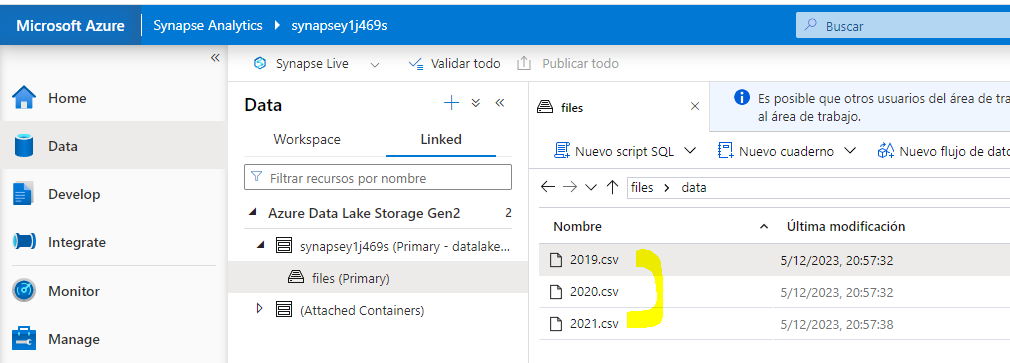
1. Expand your storage account and verify that it contains a file system container named ****files (Primary)****.



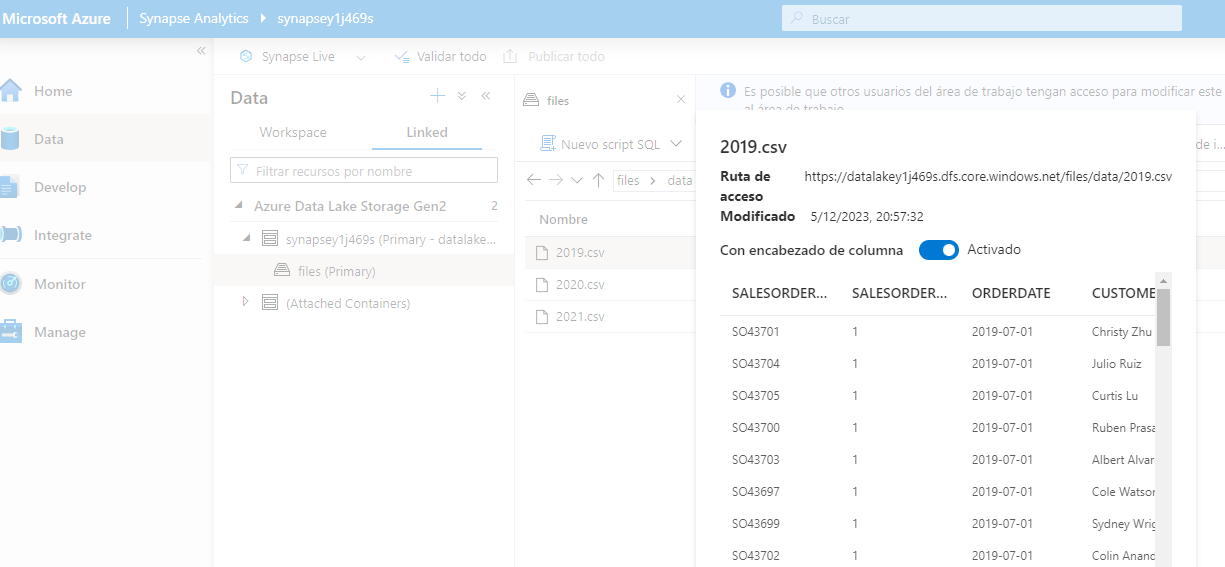
1. Select the ****files**** container, and note that it contains folders named ****data**** and ****synapse****. The synapse folder is used by Azure Synapse, and the ****data**** folder contains the data files you are going to query.



1. Open the ****data**** folder and observe that it contains .csv files for three years of sales data.



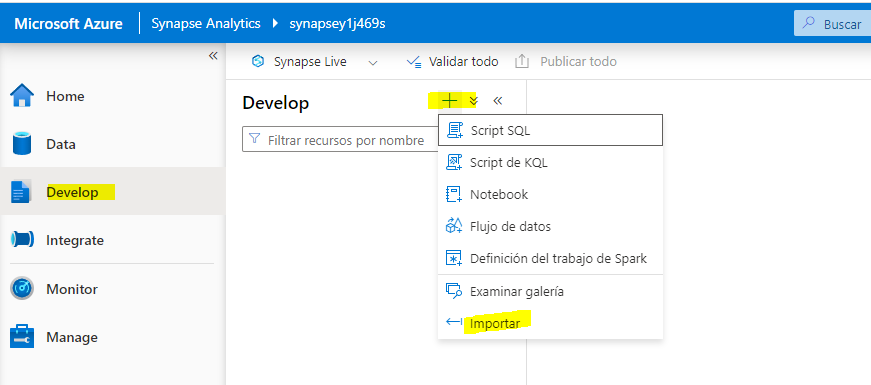
1. Right-click any of the files and select ****Preview**** to see the data it contains. Note that the files contain a header row, so you can select the option to display column headers.



1. Close the preview. Then download the ****Spark Transform.ipynb**** from [https://raw.githubusercontent.com/MicrosoftLearning/dp-203-azure-data-engineer/master/Allfiles/labs/06/notebooks/Spark%20Transform.ipynb](https://raw.githubusercontent.com/MicrosoftLearning/dp-203-azure-data-engineer/master/Allfiles/labs/06/notebooks/Spark Transform.ipynb)

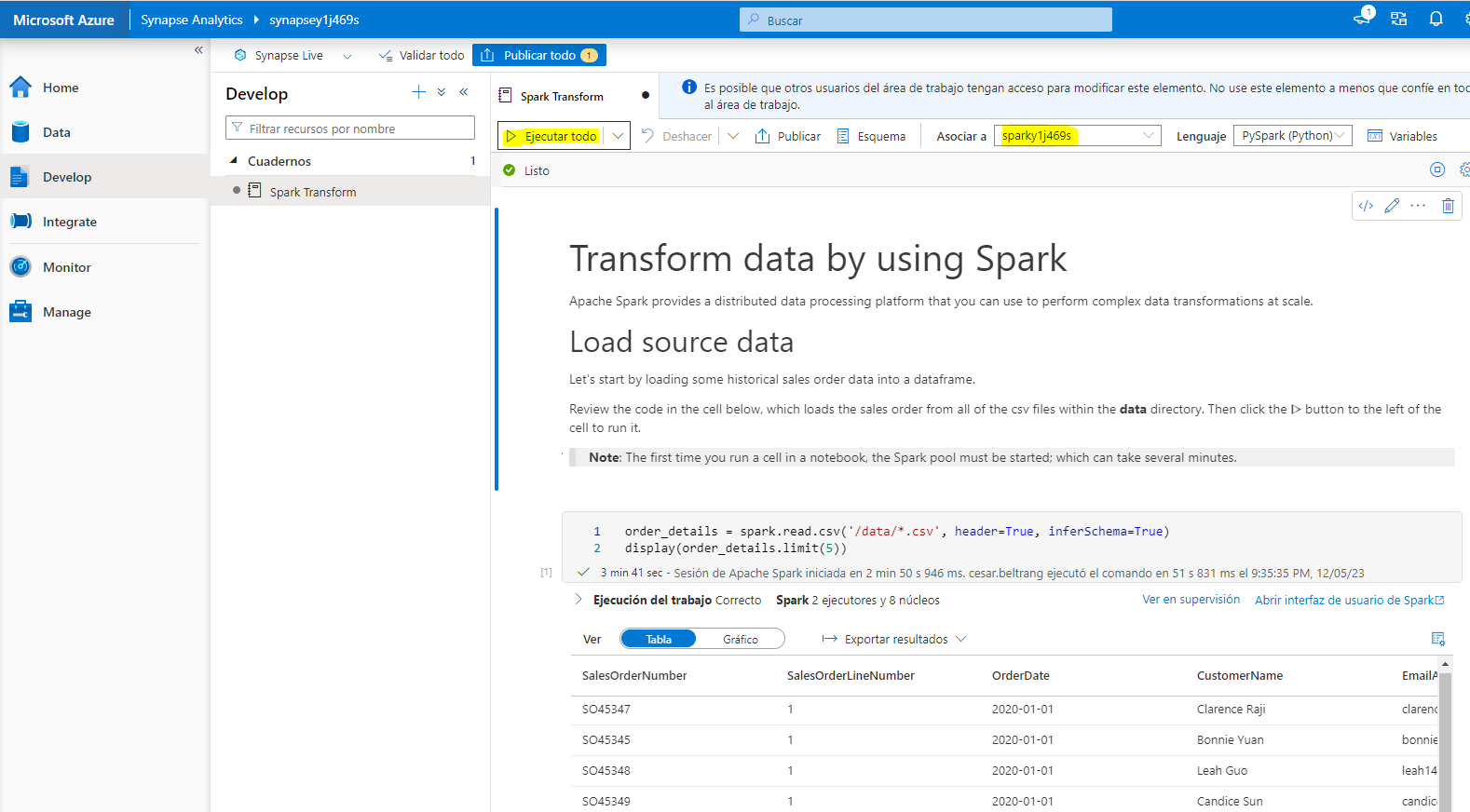
****Note****: It’s best to copy this text using the **ctrl+a** then **ctrl+c** and pasting into a tool using **ctrl+v**, such as, notepad and then using file, save as ****Spark Transform.ipynb**** with a filetype of **all files**.

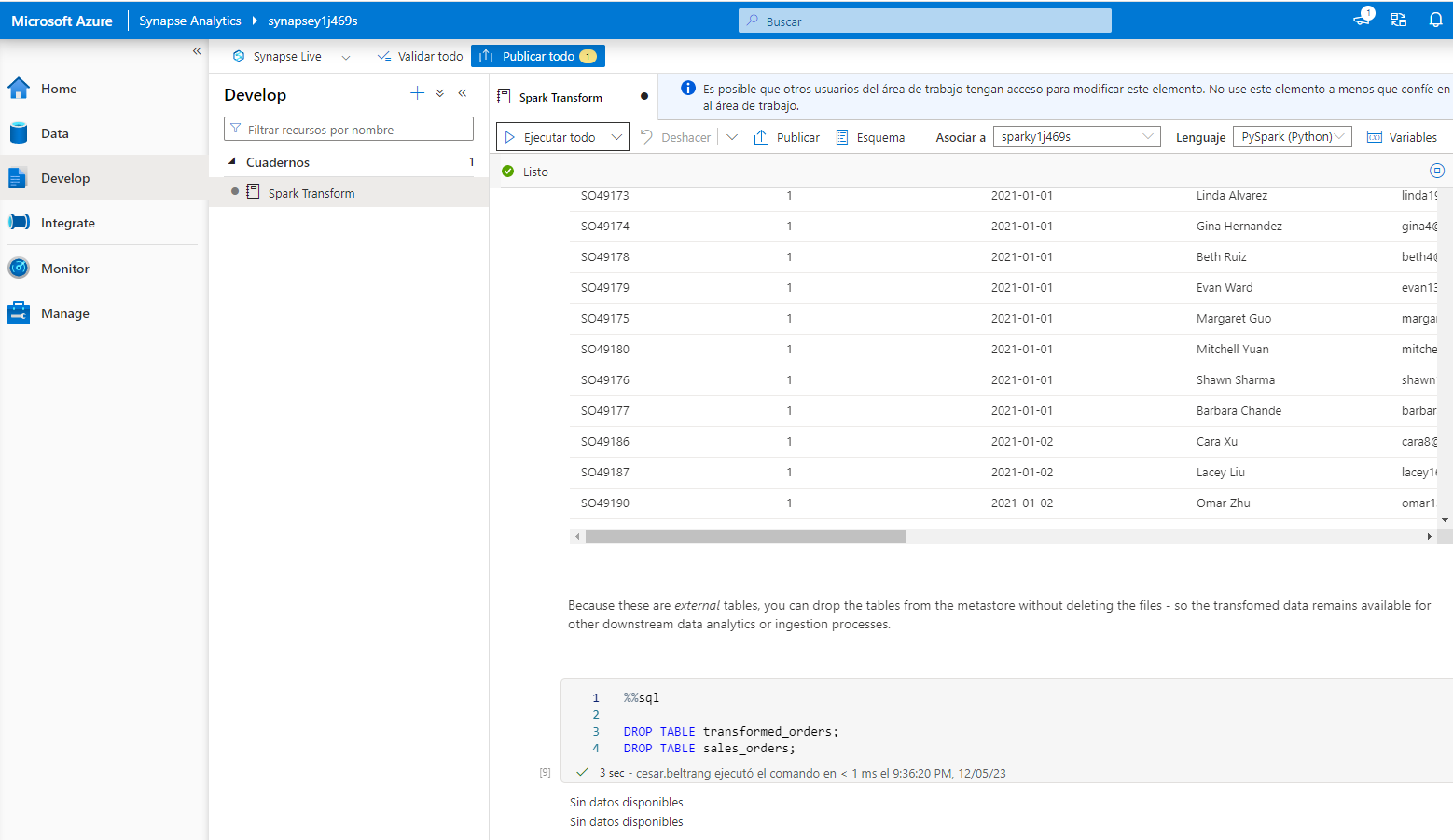
1. Then on ****Develop**** page, expand ****Notebooks**** click on the + Import options



1. Select the file you just downloaded and saved as ****Spark Transfrom.ipynb****.
2. Attach the notebook to your ****spark**xxxxxxx** Spark pool.
3. Review the notes in the notebook and run the code cells.

****Note****: The first code cell will take a few minutes to run because the Spark pool must be started. Subsequent cells will run more quickly.

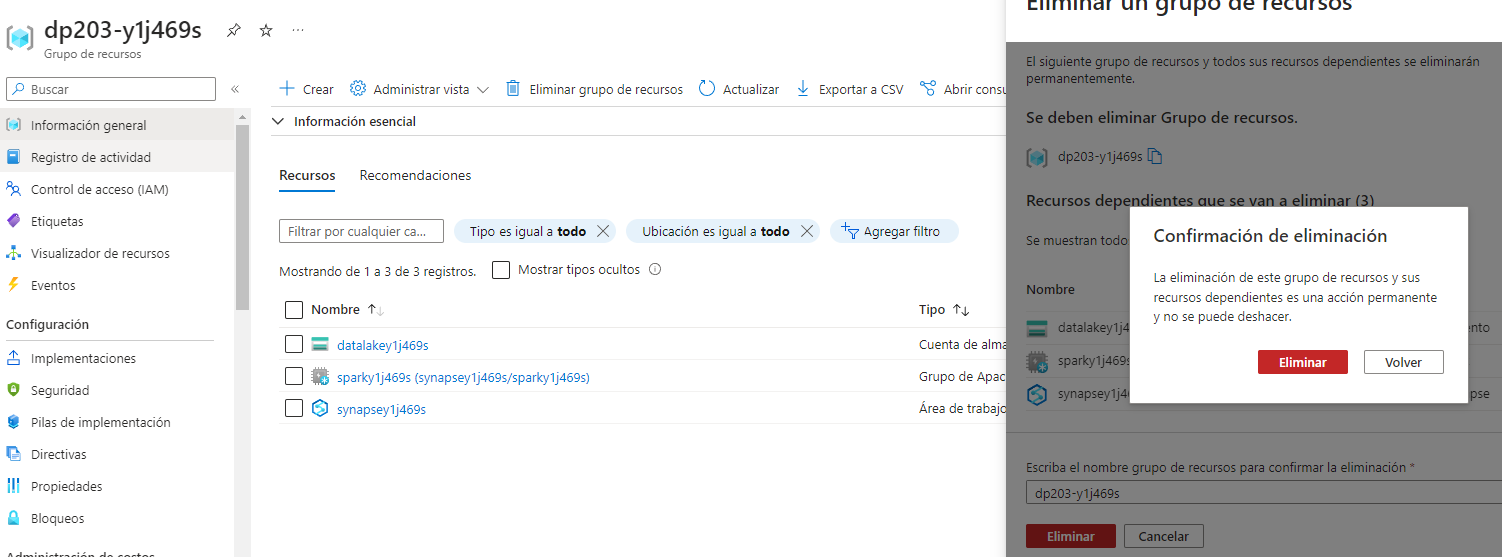




## Delete Azure resources

If you’ve finished exploring Azure Synapse Analytics, you should delete the resources you’ve created to avoid unnecessary Azure costs.

1. Close the Synapse Studio browser tab and return to the Azure portal.
2. On the Azure portal, on the ****Home**** page, select ****Resource groups****.
3. Select the ****dp203-**xxxxxxx** resource group for your Synapse Analytics workspace (not the managed resource group), and verify that it contains the Synapse workspace, storage account, and Spark pool for your workspace.
4. At the top of the ****Overview**** page for your resource group, select ****Delete resource group****.
5. Enter the ****dp203-**xxxxxxx** resource group name to confirm you want to delete it, and select ****Delete****.



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

