

# Lab 10 – Build a data pipeline in Azure Synapse Analytics

# Run a Spark notebook interactively (1)

labs.xtremelabs.io/LabViewerConnection/DetachLab...

Instance ID: 6560899  
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M11-CEP] Module 11: Use an Apache Spark notebook in a pipeline

multiple needs.

- Saves the transformed data in Parquet format in the uniquely named folder.

12. In the notebook toolbar, attach the notebook to your **sparkxxxxxxx** Spark pool and then use the ▶ **Run All** button to run all of the code cells in the notebook.

**Note:** If you find the notebook is not uploaded during the run script, you should download it from GitHub (Allfiles/labs/11/notebooks) <https://github.com/MicrosoftLearning/dp-203-azure-data-engineer/tree/master/Allfiles/labs/11/notebooks>. The file named Spark Transform.ipynb and upload it to Synapse.

The Spark session may take a few minutes to start before the code cells can run.

13. After all of the notebook cells have run, note the name of the folder in which the transformed data has been saved.

14. Switch to the **files** tab (which should still be open) and view the root files folder. If necessary, in then **More** menu, select **Refresh** to see the new folder. Then open it to verify that it contains Parquet files.

15. Return to the root **files** folder, then select the uniquely named folder generated by the notebook and in the

Page: 4/6

Support

synapse8460zn - Microsoft Azure

synapse8460zn - Azure Synapse

https://web.azuresynapse.net/en/authoring/analyz...

synapse8460zn

Synapse live

Validate all

Publish all 1

files

Spark Transform

Run all

Undo

Publish

Outline

Ready

## Transform data by using Spark

This notebook transforms sales order data; converting it from CSV to Parquet format and splitting customer name into two separate fields.

### Set variables

```
1 import uuid
2
3 # Variable for unique folder name
4 folderName = uuid.uuid4()
```

[1] ✓ 3 min 19 session started in 3 min 19 sec 500 ms. Command executed in 155 ms by XLab-P6T-900 on 1:

10°C Mostly cloudy

Search

ENG US

1:38 PM 2024-04-11

# Run a Spark notebook interactively (2)

labs.xtremelabs.io/LabViewerConnection/DetachLab...  
https://labs.xtremelabs.io/LabViewerConnection/...  
Instance ID: 6560899  
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M11-CEP] Module 11: Use an Apache Spark notebook in a pipeline

saved.

- 14. Switch to the **files** tab (which should still be open) and view the root files folder. If necessary, in then **More** menu, select **Refresh** to see the new folder. Then open it to verify that it contains Parquet files.
- 15. Return to the root **files** folder, then select the uniquely named folder generated by the notebook and in the **New SQL Script** menu, select **Select TOP 100 rows**.
- 16. In the **Select TOP 100 rows** pane, set the file type to Parquet format and apply the change.
- 17. In the new SQL Script pane that opens, use the ▶ **Run** button to run the SQL code and verify that it returns the transformed sales order data.

Page: 4/6

Support

synapse8460zn - Microsoft Azure x synapse8460zn - Azure Synapse x  
https://web.azuresynapse.net/en/authoring/analyz...  
synapse8460zn

Synapse live Validate all Publish all 2

files Spark Transform SQL script 1

Run Undo Publish Query plan Connect to Built-in

```
2 SELECT
3   TOP 100 *
4 FROM
5   OPENROWSET (Follow link (ctrl + click))
6   BULK 'https://datalake8460zn.dfs.core.windows.net/files/d55b6167-3a42-4594-8a72-
7   FORMAT = 'PARQUET'
8   AS [result]
```

Results Messages

View Table Chart Export results

SalesOrderNu...	SalesOrderLine...	OrderDate	EmailAddress	Item	Quantity
SO43701	1	2019-07-01	christy12@adv...	Mountain-100 ...	1
SO43704	1	2019-07-01	julio1@adventu...	Mountain-100 ...	1
SO43705	1	2019-07-01	curtis9@advent...	Mountain-100 ...	1

00:00:14 Query executed successfully.

10°C Mostly cloudy 1:42 PM 2024-04-11

Run the notebook in a pipeline

# Create a parameters cell

The image shows a dual-monitor setup. The left monitor displays a web browser at <https://labs.xtremelabs.io/LabViewerConnection/DetachLab...>. The page has a header with 'Instance ID: 6560899' and 'XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M11-CEP] Module 11: Use an Apache Spark notebook in a pipeline'. The main content area is titled 'Create a parameters cell' and contains a list of four steps, each with a blue checkmark icon:

- Now that you understand the transformation process, you're ready to automate it by encapsulating the notebook in a pipeline.
1. In Synapse Studio, return to the **Spark Transform** tab that contains the notebook, and in the toolbar, in the ... menu at the right end, select **Clear output**.
2. Select the first code cell (which contains the code to set the **folderName** variable).
3. In the pop-up toolbar at the top right of the code cell, in the ... menu, select **[@] Toggle parameter cell**. Verify that the word **parameters** appears at the bottom right of the cell.
4. In the toolbar, use the **Publish** button to save the changes.

Below this list is a section titled 'Create a pipeline' with two steps, each with a checkbox icon:

- ☐ 1. In Synapse Studio, select the **Integrate** page. Then in the + menu select **Pipeline** to create a new pipeline.
- ☐ 2. In the **Properties** pane for your new pipeline,

At the bottom of the left monitor, there is a 'Support' button and a 'Page: 5/6' indicator.

The right monitor displays the Microsoft Azure Synapse Studio interface. The browser address bar shows <https://web.azuresynapse.net/en/authoring/analyz...>. The Synapse Studio header shows 'synapseseg8460zn' and a search bar. The main workspace is titled 'Transform data by using Spark' and contains a code cell with the following Python code:

```
1 import uuid
2
3 # Variable for unique folder name
4 folderName = uuid.uuid4()
```

Below the code cell, the word 'Parameters' is visible. The bottom of the right monitor shows the Windows taskbar with various application icons and a system clock indicating 1:45 PM on 2024-04-11.

# Create a pipeline

The screenshot displays the Azure Synapse Studio interface for creating a pipeline. On the left, a sidebar provides instructions for adding a Notebook activity to a pipeline. The main workspace shows a pipeline named 'Transform Sales D...' with a 'Notebook' activity selected. The 'Settings' tab for the Notebook activity is open, showing configuration options like 'sparkg8460zn' for the pool and 'Small(4 vCores, 28GB memory)' for the executor size. The bottom of the screen shows the Windows taskbar with various application icons and system status information.

drag a **Notebook** activity to the pipeline design surface containing a Notebook activity.

- 4. In the **General** tab for the Notebook activity, change its name to **Run Spark Transform**.
- 5. In the **Settings** tab for the Notebook activity, set the following properties:
  - **Notebook:** Select the **Spark Transform** notebook.
  - **Base parameters:** Expand this section and define a parameter with the following settings:
    - **Name:** `folderName`
    - **Type:** String
    - **Value:** Select Add dynamic content and set the parameter value to the *Pipeline Run ID* system variable `(@pipeline().RunId)`
  - **Spark pool:** Select the **sparkxxxxxxx** pool.
  - **Executor size:** Select Small (4 vCores, 28GB Memory).

Your pipeline containing a Notebook activity with settings.

### Publish and run the pipeline

- 1. Use the **Publish all** button to publish the pipeline (and any other unsaved assets).

Page: 5/6

Support

synapseg8460zn - Microsoft Azure Synapse

https://web.azuresynapse.net/en/authoring/orche...

synapseg8460zn

Search

Synapse live Validate all Publish all 2

es Spark Transform SQL script 1 Transform Sales D...

Activities

Search activities

Synapse

- Notebook
- Spark job definition
- SQL pool stored proc...

Move and transform

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Notebook

Run Spark Transform

General Settings User properties

sparkg8460zn

Executor size ①

Small(4 vCores, 28GB memory)

Add dynamic content [Alt+Shift+D]

Dynamically allocate executors ① ☐ Enabled ☐ Disabled

Min executors ①

10°C Cloudy

Search

ENG US

1:54 PM 2024-04-11

# Publish and run the pipeline

The image displays two side-by-side browser windows. The left window is the Xtremelabs LabViewer, showing a list of steps for running a pipeline. The right window is the Azure Synapse Studio interface, showing the 'files' storage container with a list of folders created by the pipeline run.

**Left Window (Xtremelabs LabViewer):**

- Instance ID: 6560899
- Xtremelabs: DP-203T00-A-CEP [DP-203T00-A-M11-CEP] Module 11: Use an Apache Spark notebook in a pipeline
- Step 2: At the top of the pipeline designer pane, in the **Add trigger** menu, select **Trigger now**. Then select **OK** to confirm you want to run the pipeline.
- Note: You can also create a trigger to run the pipeline at a scheduled time or in response to a specific event.
- Step 3: When the pipeline has started running, on the **Monitor** page, view the **Pipeline runs** tab and review the status of the **Transform Sales Data** pipeline.
- Step 4: Select the **Transform Sales Data** pipeline to view its details, and note the Pipeline run ID in the **Activity runs** pane.
- The pipeline may take five minutes or longer to complete. You can use the **Refresh** button on the toolbar to check its status.
- Step 5: When the pipeline run has succeeded, on the **Data** page, browse to the **files** storage container and verify that a new folder named for the pipeline run ID has been created, and that it contains Parquet files for the transformed sales data.

**Right Window (Azure Synapse Studio):**

- URL: <https://web.azuresynapse.net/en/authoring/explor...>
- Search bar: Search
- Navigation: Synapse live, Validate all, Publish all
- Files Explorer: files
- Actions: New SQL script, New data flow, New integration dataset, Upload, More
- Table of Files:

Name	Last Modified	Content Type	Size
47dd9296-14c7-4eb3-87bb-acd68555c8af	2024-04-11, 2:00:21 p.m.	Folder	
d55b6167-3a42-4594-8a72-47758d72aa7f	2024-04-11, 1:33:30 p.m.	Folder	
data	2024-04-11, 1:21:53 p.m.	Folder	
synapse	2024-04-11, 1:21:13 p.m.	Folder	

**Bottom Window (File Explorer):**

- Path: Data > 47dd
- File Name: 47dd9296-14c7-4eb3-87bb-acd68555c8af

**Page 5/6**