USE DELTA LAKE IN AZURE SYNAPSE ANALYTICS

BY NITIN

OBJECTIVES

- Understood how to use Delta Lake with Spark in Azure Synapse Analytics.
- I learned about how to create external tables and manage tables.
- Understood how to load the data into a Delta Table object.

labs.xtremelabs.io/LabViewerConnection/DetachLabManu...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?la...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

4. Use Delta Lake with Spark in Azure Synapse Analytics

Create delta tables:

- [x] The script provisions an Azure Synapse Analytics workspace and an Azure Storage account to host the data lake, then uploads a data file to the data lake.

Explore the data in the data lake

- [x] 1. After the script has completed, in the Azure portal, go to the **dp203-xxxxxxx** resource group that it created, and select your Synapse workspace.
- [x] 2. In the Overview page for your Synapse workspace, in the **Open Synapse Studio** card, select **Open** to open Synapse Studio in a new browser tab; signing in if prompted.
- [x] 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 **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 **synapsexxxxxxx (Primary - datalakexxxxxxx)**.
- [ ] 5. Expand your storage account and verify that it contains a file system container named **files**.
- [ ] 6. Select the **files** container, and note that it contains a folder named **products**. This folder contains the data you are going to work with in this exercise.
- [ ] 7. Open the **products** folder, and observe that it contains a file named **products.csv**.
- [ ] 8. Select **products.csv**, and then in the **New notebook** list on the toolbar, select **Load to DataFrame**.
- [ ] 9. In the **Notebook 1** pane that opens, in the **Attach** to list, select the **sparkxxxxxxx** Spark pool and ensure that the **Language** is set to **PySpark (Python)**.

Page: 4/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked?workspace=%2Fsubscriptions%2F9353cc25-ed82-417d-9...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Home
Data
Develop
Integrate
Monitor
Manage

Synapse live | Validate all | Publish all

Data

Workspace | Linked

Filter resources by name

Azure Data Lake Storage Gen2 2

synapseqyr3wu5 (Primary - datalak...

(Attached Containers)

Select an item

Use the resource explorer to select or create a new item

Search

ENG IN

6:44 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManu...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?la...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

```
%%pyspark
df = spark.read.load('abfss://files@datalakexxxxxxx.dfs.core.
windows.net/products/products.csv', format='csv'
## If header exists uncomment line below
##, header=True
)
display(df.limit(10))
```

11. Uncomment the , *header=True* line (because the products.csv file has the column headers in the first line), so your code looks like this:

Paste Content

Paste Content

```
%%pyspark
df = spark.read.load('abfss://files@datalakexxxxxxx.dfs.core.
windows.net/products/products.csv', format='csv'
## If header exists uncomment line below
, header=True
)
display(df.limit(10))
```

12. Use the ▷ icon to the left of the code cell to run it, and wait for the results. The first time you run a cell in a notebook, the Spark pool is started - so it may take a minute or so to return any results. Eventually, the results should appear below the cell, and they should be similar to this:

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3399.9900
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
...	...	...	...

Previous Exercise Next Exercise

Page: 4/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscription...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all 1

files | Notebook 1 ●

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run all | Undo | Publish | Outline | Attach to sparkqyr3wu5 | Language PySpark (Python) | Variables

Ready

```
1  %%pyspark
2  df = spark.read.load('abfss://files@datalakeqyr3wu5.dfs.core.windows.net/products/products.csv', format='csv'
3  ## If header exists uncomment line below
4  ##, header=True
5  )
6  display(df.limit(10))
```

[1] ✓ 4 min 14 sec - Apache Spark session started in 3 min 43 sec 208 ms. Command executed in 31 sec 416 ms by XLab-k8V-976 on 6:50:51 PM, 11/13/24

Job execution Succeeded **Spark** 2 executors 8 cores View in monitoring Open Spark UI

View: Table | Chart Export results

_c0	_c1	_c2	_c3
ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3399.9900
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
773	Mountain-100 Silver, 44	Mountain Bikes	3399.9900
774	Mountain-100 Silver, 48	Mountain Bikes	3399.9900
775	Mountain-100 Black, 38	Mountain Bikes	3374.9900
776	Mountain-100 Black, 42	Mountain Bikes	3374.9900
777	Mountain-100 Black, 44	Mountain Bikes	3374.9900
778	Mountain-100 Black, 48	Mountain Bikes	3374.9900
779	Mountain-200 Silver, 38	Mountain Bikes	2319.9900

Search | ENG IN | 6:50 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

```
)
display(df.limit(10))
```

11. Uncommment the , *header=True* line (because the products.csv file has the column headers in the first line), so your code looks like this:
Paste Content
Paste Content

```
%%pyspark
df = spark.read.load('abfss://files@datalakexxxxxxx.dfs.core.windows.net/products/products.csv', format='csv'
## If header exists uncomment line below
, header=True
)
display(df.limit(10))
```

12. Use the ▷ icon to the left of the code cell to run it, and wait for the results. The first time you run a cell in a notebook, the Spark pool is started - so it may take a minute or so to return any results. Eventually, the results should appear below the cell, and they should be similar to this:

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3399.9900
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
...	...	...	...

Previous Exercise Next Exercise

Page: 4/10

Support

XtremeLabs - [DP-203T00-A-M synapseqyr3wu5 - Microsoft Az synapseqyr3wu5 - Azure Synap

web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscri...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 Search XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live Validate all Publish all 1

files Notebook 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run all Undo Publish Outline Attach to sparkqyr3wu5 Language PySpark (Python) Variables

Ready

```
1  %%pyspark
2  df = spark.read.load('abfss://files@datalakeqyr3wu5.dfs.core.windows.net/products/products.csv', format='csv'
3  ## If header exists uncomment line below
4  , header=True
5  )
6  display(df.limit(10))
```

[2] 5 sec - Command executed in 5 sec 348 ms by XLab-k8V-976 on 6:52:49 PM, 11/13/24

Job execution Succeeded Spark 2 executors 8 cores View in monitoring Open Spark UI

View: Table | Chart Export results

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3399.9900
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
773	Mountain-100 Silver, 44	Mountain Bikes	3399.9900
774	Mountain-100 Silver, 48	Mountain Bikes	3399.9900
775	Mountain-100 Black, 38	Mountain Bikes	3374.9900
776	Mountain-100 Black, 42	Mountain Bikes	3374.9900
777	Mountain-100 Black, 44	Mountain Bikes	3374.9900
778	Mountain-100 Black, 48	Mountain Bikes	3374.9900
779	Mountain-200 Silver, 38	Mountain Bikes	2319.9900
780	Mountain-200 Silver, 42	Mountain Bikes	2319.9900

Search ENG IN 6:53 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

41% Completed

Lab Time Left: 01:34:09

Lab Actions

5. Use Delta Lake with Spark in Azure Synapse Analytics

Load the file data into a delta table:

1. Under the results returned by the first code cell, use the + **Code** button to add a new code cell. Then enter the following code in the new cell and run it:

Paste Content
Paste Content

```
delta_table_path = "/delta/products-delta"
df.write.format("delta").save(delta_table_path)
```

2. On the **files** tab, use the ↑ icon in the toolbar to return to the root of the files container, and note that a new folder named **delta** has been created. Open this folder and the **products-delta** table it contains, where you should see the parquet format file(s) containing the data.

3. Return to the Notebook 1 tab and add another new code cell. Then, in the new cell, add the following code and run it:

Paste Content
Paste Content

```
from delta.tables import *
from pyspark.sql.functions import *

Create a deltaTable object
deltaTable = DeltaTable.forPath(spark, delta_table_path)

Update the table (reduce price of product 771 by 10%)
deltaTable.update(
    condition = "ProductID == 771",
    set = { "ListPrice": "ListPrice * 0.9" })

View the updated data as a dataframe
```

Page: 5/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/storageaccounts/synapseqyr3wu5-WorkspaceDefa...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all 1

files | Notebook 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

New SQL script | New data flow | New integration dataset | Upload | Download | New folder | Select all | More

files

Name	Last Modified	Content Type	Size
delta	11/13/2024, 6:54:25 PM	Folder	
products	11/13/2024, 6:42:59 PM	Folder	
synapse	11/13/2024, 6:42:19 PM	Folder	

Showing 1 to 3 of 3 cached items

Search

ENG IN

6:55 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

Paste Content

```
from delta.tables import *
from pyspark.sql.functions import *

Create a deltaTable object
deltaTable = DeltaTable.forPath(spark, delta_table_path)

Update the table (reduce price of product 771 by 10%)
deltaTable.update(
    condition = "ProductID == 771",
    set = { "ListPrice": "ListPrice * 0.9" })

View the updated data as a dataframe
deltaTable.toDF().show(10)
```

The data is loaded into a **DeltaTable** object and updated. You can see the update reflected in the query results.

4. Add another new code cell with the following code and run it:

Paste Content
Paste Content

```
new_df = spark.read.format("delta").load(delta_table_path)
new_df.show(10)
```

The code loads the delta table data into a data frame from its location in the data lake, verifying that the change you made via a **DeltaTable** object has been persisted.

5. Modify the code you just ran as follows, specifying the option to use the time travel feature of delta lake to view a previous version of the data.

Paste Content
Paste Content

```
new_df = spark.read.format("delta").option("versionAsOf", 0).load(d
elta_table_path)
new_df.show(10)
```

When you run the modified code, the results show the original version of the data.

6. Add another new code cell with the following code and run it:

Paste Content

Page: 5/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscri...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all 1

files | Notebook 1 ●

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run all | Undo | Publish | Outline | Attach to sparkqyr3wu5 | Language PySpark (Python) | Variables

Ready

```
1  new_df = spark.read.format("delta").load(delta_table_path)
2  new_df.show(10)
```

[6] ✓ 1 sec - Command executed in 1 sec 73 ms by XLab-k8V-976 on 6:59:59 PM, 11/13/24

Job execution Succeeded **Spark** 2 executors 8 cores View in monitoring Open Spark UI

```
+---------+--------------------+--------------+---------+
|ProductID|         ProductName|      Category|ListPrice|
+---------+--------------------+--------------+---------+
|      771|Mountain-100 Silv...|Mountain Bikes| 3059.991|
|      772|Mountain-100 Silv...|Mountain Bikes|3399.9900|
|      773|Mountain-100 Silv...|Mountain Bikes|3399.9900|
|      774|Mountain-100 Silv...|Mountain Bikes|3399.9900|
|      775|Mountain-100 Blac...|Mountain Bikes|3374.9900|
|      776|Mountain-100 Blac...|Mountain Bikes|3374.9900|
|      777|Mountain-100 Blac...|Mountain Bikes|3374.9900|
|      778|Mountain-100 Blac...|Mountain Bikes|3374.9900|
|      779|Mountain-200 Silv...|Mountain Bikes|2319.9900|
|      780|Mountain-200 Silv...|Mountain Bikes|2319.9900|
+---------+--------------------+--------------+---------+
only showing top 10 rows
```

Search | ENG IN | 7:00 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

The data is loaded into a **DeltaTable** object and updated. You can see the update reflected in the query results.

4. Add another new code cell with the following code and run it:

Paste Content
Paste Content

```
new_df = spark.read.format("delta").load(delta_table_path)
new_df.show(10)
```

The code loads the delta table data into a data frame from its location in the data lake, verifying that the change you made via a **DeltaTable** object has been persisted.

5. Modify the code you just ran as follows, specifying the option to use the time travel feature of delta lake to view a previous version of the data.

Paste Content
Paste Content

```
new_df = spark.read.format("delta").option("versionAsOf", 0).load(delta_table_path)
new_df.show(10)
```

When you run the modified code, the results show the original version of the data.

6. Add another new code cell with the following code and run it:

Paste Content
Paste Content

```
deltaTable.history(10).show(20, False, True)
```

The history of the last 20 changes to the table is shown - there should be two (the original creation, and the update you made.)

Previous Exercise Next Exercise

Page: 5/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscri...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all 1

files | Notebook 1 ●

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run all | Undo | Publish | Outline | Attach to sparkqyr3wu5 | Language PySpark (Python) | Variables

Ready

```
1  deltaTable.history(10).show(20, False, True)
```

[8] 3 sec - Command executed in 2 sec 805 ms by XLab-k8V-976 on 7:01:46 PM, 11/13/24

Job execution Succeeded **Spark** 2 executors 8 cores View in monitoring Open Spark UI

```
-RECORD 0-----------------------------------------------------------------------------------------------------------------------------------------------------
 version             | 1
 timestamp           | 2024-11-14 01:58:48.922
 userId              | null
 userName            | null
 operation           | UPDATE
 operationParameters | {predicate -> ["(cast(ProductID#225 as int) = 771)"]}
 job                 | null
 notebook            | null
 clusterId           | null
 readVersion         | 0
 isolationLevel      | Serializable
 isBlindAppend       | false
 operationMetrics    | {numRemovedFiles -> 1, numRemovedBytes -> 6233, numCopiedRows -> 294, numAddedChangeFiles -> 0, executionTimeMs ->
28244, scanTimeMs -> 25598, numAddedFiles -> 1, numUpdatedRows -> 1, numAddedBytes -> 6245, rewriteTimeMs -> 2637}
 userMetadata        | null
 engineInfo          | Apache-Spark/3.4.3.5.3.20241016.1 Delta-Lake/2.4.0.19
-RECORD 1-----------------------------------------------------------------------------------------------------------------------------------------------------
 version             | 0
 timestamp           | 2024-11-14 01:54:33.967
 userId              | null
 userName            | null
 operation           | WRITE
 operationParameters | {mode -> ErrorIfExists, partitionBy -> []}
 job                 | null
 notebook            | null
```

Search | ENG IN | 7:02 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

- *Managed* tables, that are defined in the Hive metastore for the Spark pool.

Create an external table

1. In a new code cell, add and run the following code:

Paste Content
Paste Content

```
spark.sql("CREATE DATABASE AdventureWorks")
spark.sql("CREATE TABLE AdventureWorks.ProductsExternal USING DELTA
LOCATION '{0}'".format(delta_table_path))
spark.sql("DESCRIBE EXTENDED AdventureWorks.ProductsExternal").show
(truncate=False)
```

This code creates a new database named **AdventureWorks** and then creates an external tabled named **ProductsExternal** in that database based on the path to the parquet files you defined previously. It then displays a description of the table's properties. Note that the **Location** property is the path you specified.

2. Add a new code cell, and then enter and run the following code:

Paste Content
Paste Content

```
%%sql

USE AdventureWorks;

SELECT * FROM ProductsExternal;
```

The code uses SQL to switch context to the **AdventureWorks** database (which returns no data) and then query the **ProductsExternal** table (which returns a resultset containing the products data in the Delta Lake table).

Previous Exercise Next Exercise

Page: 6/10

Support

XtremeLabs - [DP-203T00-A-M synapseqyr3wu5 - Microsoft Az synapseqyr3wu5 - Azure Synap

web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscri...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 Search XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live Validate all Publish all 1

files Notebook 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run all Undo Publish Outline Attach to sparkqyr3wu5 Language PySpark (Python) Variables

Ready

```
1  spark.sql("CREATE DATABASE AdventureWorks")
2  spark.sql("CREATE TABLE AdventureWorks.ProductsExternal USING DELTA LOCATION '{0}'".format(delta_table_path))
3  spark.sql("DESCRIBE EXTENDED AdventureWorks.ProductsExternal").show(truncate=False)
```

[9] 16 sec - Command executed in 16 sec 807 ms by XLab-k8V-976 on 7:03:01 PM, 11/13/24

```
+----------------------------+----------------------------------------------------------------------------+-------+
|col_name                    |data_type                                                                   |comment|
+----------------------------+----------------------------------------------------------------------------+-------+
|ProductID                   |string                                                                      |null   |
|ProductName                 |string                                                                      |null   |
|Category                    |string                                                                      |null   |
|ListPrice                   |string                                                                      |null   |
|                            |                                                                            |       |
|# Detailed Table Information|                                                                            |       |
|Name                        |spark_catalog.adventureworks.productsexternal                               |       |
|Type                        |EXTERNAL                                                                    |       |
|Location                    |abfss://files@datalakeqyr3wu5.dfs.core.windows.net/delta/products-delta|       |
|Provider                    |delta                                                                       |       |
|Owner                       |trusted-service-user                                                        |       |
|Table Properties            |[delta.minReaderVersion=1,delta.minWriterVersion=2]                         |       |
+----------------------------+----------------------------------------------------------------------------+-------+
```

Search ENG IN 7:03 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

- *Managed* tables, that are defined in the Hive metastore for the Spark pool.

Create an external table

1. In a new code cell, add and run the following code:

Paste Content
Paste Content

```
spark.sql("CREATE DATABASE AdventureWorks")
spark.sql("CREATE TABLE AdventureWorks.ProductsExternal USING DELTA
LOCATION '{0}'".format(delta_table_path))
spark.sql("DESCRIBE EXTENDED AdventureWorks.ProductsExternal").show
(truncate=False)
```

This code creates a new database named **AdventureWorks** and then creates an external tabled named **ProductsExternal** in that database based on the path to the parquet files you defined previously. It then displays a description of the table's properties. Note that the **Location** property is the path you specified.

2. Add a new code cell, and then enter and run the following code:

Paste Content
Paste Content

```
%%sql

USE AdventureWorks;

SELECT * FROM ProductsExternal;
```

The code uses SQL to switch context to the **AdventureWorks** database (which returns no data) and then query the **ProductsExternal** table (which returns a resultset containing the products data in the Delta Lake table).

Previous Exercise Next Exercise

Page: 6/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscri...

Microsoft Azure | Synapse Analytics | synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all 1

files | Notebook 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run all | Undo | Publish | Outline | Attach to sparkqyr3wu5 | Language PySpark (Python) | Variables

Ready

```
1  %%sql
2
3  USE AdventureWorks;
4
5  SELECT * FROM ProductsExternal;
```

[10] 5 sec - Command executed in 4 sec 842 ms by XLab-k8V-976 on 7:04:11 PM, 11/13/24

Job execution Succeeded **Spark** 2 executors 8 cores View in monitoring Open Spark UI

No data available

View: Table | Chart Export results

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3059.991
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
773	Mountain-100 Silver, 44	Mountain Bikes	3399.9900
774	Mountain-100 Silver, 48	Mountain Bikes	3399.9900
775	Mountain-100 Black, 38	Mountain Bikes	3374.9900
776	Mountain-100 Black, 42	Mountain Bikes	3374.9900
777	Mountain-100 Black, 44	Mountain Bikes	3374.9900
778	Mountain-100 Black, 48	Mountain Bikes	3374.9900
779	Mountain-200 Silver, 38	Mountain Bikes	2319.9900
780	Mountain-200 Silver, 42	Mountain Bikes	2319.9900
781	Mountain-200 Silver, 46	Mountain Bikes	2319.9900
782	Mountain-200 Black, 38	Mountain Bikes	2294.9900
783	Mountain-200 Black, 42	Mountain Bikes	2294.9900

Search ENG IN 7:04 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

Lab Time Left: 01:23:18

Lab Actions

7. Use Delta Lake with Spark in Azure Synapse Analytics

Create a managed table:

1. In a new code cell, add and run the following code:

Paste Content
Paste Content

```
df.write.format("delta").saveAsTable("AdventureWorks.ProductsManaged")
spark.sql("DESCRIBE EXTENDED AdventureWorks.ProductsManaged").show(truncate=False)
```

This code creates a managed tabled named **ProductsManaged** based on the DataFrame you originally loaded from the **products.csv** file (before you updated the price of product 771). You do not specify a path for the parquet files used by the table - this is managed for you in the Hive metastore, and shown in the **Location** property in the table description (in the **files/synapse/workspaces/synapsexxxxxxx/warehouse** path).

2. Add a new code cell, and then enter and run the following code:

Paste Content
Paste Content

```
%%sql

USE AdventureWorks;

SELECT * FROM ProductsManaged;
```

The code uses SQL to query the **ProductsManaged** table.

Compare external and managed tables

1. In a new code cell, add and run the following code:

Paste Content
Paste Content

```
%%sql
```

Page: 7/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscri...

Microsoft Azure | Synapse Analytics | synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all 1

files | Notebook 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run all | Undo | Publish | Outline | Attach to sparkqyr3wu5 | Language PySpark (Python) | Variables

Ready

```
1  %%sql
2
3  USE AdventureWorks;
4
5  SELECT * FROM ProductsManaged;
```

[12] 8 sec - Command executed in 6 sec 923 ms by XLab-k8V-976 on 7:05:50 PM, 11/13/24

Job execution Succeeded **Spark** 2 executors 8 cores | View in monitoring | Open Spark UI

No data available

View: Table | Chart | Export results

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3399.9900
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
773	Mountain-100 Silver, 44	Mountain Bikes	3399.9900
774	Mountain-100 Silver, 48	Mountain Bikes	3399.9900
775	Mountain-100 Black, 38	Mountain Bikes	3374.9900
776	Mountain-100 Black, 42	Mountain Bikes	3374.9900
777	Mountain-100 Black, 44	Mountain Bikes	3374.9900
778	Mountain-100 Black, 48	Mountain Bikes	3374.9900
779	Mountain-200 Silver, 38	Mountain Bikes	2319.9900
780	Mountain-200 Silver, 42	Mountain Bikes	2319.9900
781	Mountain-200 Silver, 46	Mountain Bikes	2319.9900
782	Mountain-200 Black, 38	Mountain Bikes	2294.9900
783	Mountain-200 Black, 42	Mountain Bikes	2294.9900

Search

ENG IN | 7:05 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

```
# Write the stream to a delta table
delta_stream_table_path = '/delta/iotdevicedata'
checkpointpath = '/delta/checkpoint'
deltastream = iotstream.writeStream.format("delta").option("checkpo
intLocation", checkpointpath).start(delta_stream_table_path)
print("Streaming to delta sink...")
```

This code writes the streaming device data in delta format.

- [x] 3. In a new code cell, add and run the following code:

Paste Content
Paste Content

```
# Read the data in delta format into a dataframe
df = spark.read.format("delta").load(delta_stream_table_path)
display(df)
```

This code reads the streamed data in delta format into a dataframe. Note that the code to load streaming data is no different to that used to load static data from a delta folder.

- [ ] 4. In a new code cell, add and run the following code:

Paste Content
Paste Content

```
# create a catalog table based on the streaming sink
spark.sql("CREATE TABLE IotDeviceData USING DELTA LOCATION '{0}'".f
ormat(delta_stream_table_path))
```

This code creates a catalog table named **IotDeviceData** (in the **default** database) based on the delta folder. Again, this code is the same as would be used for non-streaming data.

- [ ] 5. In a new code cell, add and run the following code:

Paste Content
Paste Content

```
%%sql

SELECT * FROM IotDeviceData;
```

This code queries the **IotDeviceData** table, which contains the device data from the streaming source.

- [ ] 6. In a new code cell, add and run the following code:

Page: 8/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscri...

Microsoft Azure | Synapse Analytics | synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all 1

files | Notebook 1 ●

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run all | Undo | Publish | Outline | Attach to sparkqyr3wu5 | Language PySpark (Python) | Variables

Ready

```
Streaming to delta sink...
```

```
1  # Read the data in delta format into a dataframe
2  df = spark.read.format("delta").load(delta_stream_table_path)
3  display(df)
```

[20] ✓ 7 sec - Command executed in 6 sec 868 ms by XLab-k8V-976 on 7:16:17 PM, 11/13/24

Job execution Succeeded **Spark** 2 executors 8 cores | View in monitoring | Open Spark UI

View: Table | Chart | Export results

device	status
Dev1	ok
Dev1	ok
Dev1	ok
Dev2	error
Dev1	ok
Dev1	error
Dev2	ok
Dev2	error
Dev1	ok

Search | ENG IN | 7:16 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

Paste Content
Paste Content

```
# Add more data to the source stream
more_data = '''{"device":"Dev1","status":"ok"}
{"device":"Dev1","status":"ok"}
{"device":"Dev1","status":"ok"}
{"device":"Dev1","status":"ok"}
{"device":"Dev1","status":"error"}
{"device":"Dev2","status":"error"}
{"device":"Dev1","status":"ok"}'''

mssparkutils.fs.put(inputPath + "more-data.txt", more_data, True)
```

This code writes more hypothetical device data to the streaming source.

- [x] 7. In a new code cell, add and run the following code:

Paste Content
Paste Content

```
%%sql

SELECT * FROM IotDeviceData;
```

This code queries the **IotDeviceData** table again, which should now include the additional data that was added to the streaming source.

- [ ] 8. In a new code cell, add and run the following code:

Paste Content
Paste Content

```
deltastream.stop()
```

This code stops the stream.

Previous Exercise Next Exercise

Page: 8/10

Support

XtremeLabs - [DP-203T00-A-M synapseqyr3wu5 - Microsoft Az synapseqyr3wu5 - Azure Synap

web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscri...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 Search XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live Validate all Publish all 1

files Notebook 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run all Undo Publish Outline Attach to sparkqyr3wu5 Language PySpark (Python) Variables

Ready

```
1  %%sql
2
3  SELECT * FROM IotDeviceData;
```

[24] 7 sec - Command executed in 6 sec 856 ms by XLab-k8V-976 on 7:18:10 PM, 11/13/24

Job execution Succeeded **Spark** 2 executors 8 cores View in monitoring Open Spark UI

View: Table | Chart Export results

device	status
Dev1	ok
Dev1	ok
Dev1	ok
Dev2	error
Dev1	ok
Dev1	error
Dev2	ok
Dev2	error
Dev1	ok
Dev1	ok
Dev1	ok
Dev1	ok
Dev1	ok
Dev1	error

Search ENG IN 7:18 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

80% Completed

Lab Time Left: 01:09:03

Lab Actions

9. Use Delta Lake with Spark in Azure Synapse Analytics

Query a delta table from a serverless SQL pool:

- [x] In addition to Spark pools, Azure Synapse Analytics includes a built-in serverless SQL pool. You can use the relational database engine in this pool to query delta tables using SQL.
- [ ] 1. In the **files** tab, browse to the **files/delta** folder.
- [ ] 2. Select the **products-delta** folder, and on the toolbar, in the **New SQL script** drop-down list, select **Select TOP 100 rows**.
- [ ] 3. In the **Select TOP 100 rows** pane, in the **File type** list, select **Delta format** and then select **Apply**.
- [ ] 4. Review the SQL code that is generated, which should look like this:

Paste Content
Paste Content

```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://datalakexxxxxxx.dfs.core.windows.net/files/delta/products-delta/',
        FORMAT = 'DELTA'
    ) AS [result]
```

- [ ] 5. Use the ▷ **Run** icon to run the script, and review the results. They should look similar to this:

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3059.991

Page: 9/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/storageaccounts/synapseqyr3wu5-WorkspaceDefa...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all

files | Notebook 1 | Other users in your workspa... may have access to the work...

New SQL script | New data flow | New integration dataset | Upload

files › delta › products-delta

Name

_delta_log

part-00000-e4aa42a0-3f8b-4a68-9849-755a02bb5d2f-c000.snappy.parquet

part-00000-e96fc42f-f5ef-4ae3-b111-ac43711f5b95-c000.snappy.parquet

Showing 1 to 3 of 3 cached items

Select TOP 100 rows

_delta_log

Source folder format settings
Specify the format and layout of your data.

Folder path
https://datalakeqyr3wu5.dfs.core.windows.net/files/delta/products-delta/_delta_log/

File type
Delta format

Apply | Cancel

Search

ENG IN

7:20 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

80% Completed

Lab Time Left: 01:09:03

Lab Actions

9. Use Delta Lake with Spark in Azure Synapse Analytics

Query a delta table from a serverless SQL pool:

- [x] In addition to Spark pools, Azure Synapse Analytics includes a built-in serverless SQL pool. You can use the relational database engine in this pool to query delta tables using SQL.
- [ ] 1. In the **files** tab, browse to the **files/delta** folder.
- [ ] 2. Select the **products-delta** folder, and on the toolbar, in the **New SQL script** drop-down list, select **Select TOP 100 rows**.
- [ ] 3. In the **Select TOP 100 rows** pane, in the **File type** list, select **Delta format** and then select **Apply**.
- [ ] 4. Review the SQL code that is generated, which should look like this:

Paste Content
Paste Content

```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://datalakexxxxxxx.dfs.core.windows.net/files/delta/products-delta/',
        FORMAT = 'DELTA'
    ) AS [result]
```

- [ ] 5. Use the ▷ **Run** icon to run the script, and review the results. They should look similar to this:

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3059.991

Page: 9/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/storageaccounts/synapseqyr3wu5-WorkspaceDefa...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all

files | Notebook 1 | Other users in your workspa... may have access to the work...

New SQL script | New data flow | New integration dataset | Upload

files › delta › products-delta

Name

_delta_log

part-00000-e4aa42a0-3f8b-4a68-9849-755a02bb5d2f-c000.snappy.parquet

part-00000-e96fc42f-f5ef-4ae3-b111-ac43711f5b95-c000.snappy.parquet

Showing 1 to 3 of 3 cached items

Select TOP 100 rows

_delta_log

Source folder format settings
Specify the format and layout of your data.

Folder path
https://datalakeqyr3wu5.dfs.core.windows.net/files/delta/products-delta/_delta_log/

File type
Delta format

Apply | Cancel

Search

ENG IN

7:20 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

Paste Content

```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://datalakexxxxxxx.dfs.core.windows.net/files/delta/products-delta/',
        FORMAT = 'DELTA'
    ) AS [result]
```

5. Use the ▷ **Run** icon to run the script, and review the results. They should look similar to this:

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3059.991
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
...	...	...	...

This demonstrates how you can use a serverless SQL pool to query delta format files that were created using Spark, and use the results for reporting or analysis.

6. Replace the query with the following SQL code:

Paste Content

Paste Content

```
USE AdventureWorks;

SELECT * FROM Products;
```

7. Run the code and observe that you can also use the serverless SQL pool to query Delta Lake data in catalog tables that are defined the Spark metastore.

Page: 9/10

Support

XtremeLabs - [DP-203T00-A-M... | synapseqyr3wu5 - Microsoft Az... | synapseqyr3wu5 - Azure Synap...

web.azuresynapse.net/en/authoring/explore/linked/sqlscripts/SQL%20script%201?workspace=%2Fsubs...

Microsoft Azure | Synapse Analytics ▸ synapseqyr3wu5 | Search | XLab-k8V-976@xtremelabs.us LAB DIRECTORY

Synapse live | Validate all | Publish all 2

files | Notebook 1 | SQL script 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Run | Undo | Publish | Query plan | Connect to: Built-in | Use database: master

```
1  -- This is auto-generated code
2  SELECT
3      TOP 100 *
4  FROM
5      OPENROWSET(
6          BULK 'https://datalakeqyr3wu5.dfs.core.windows.net/files/delta/products-delta/',
7          FORMAT = 'DELTA'
8      ) AS [result]
9
```

Properties

General | Related (0)

Name * SQL script 1

Description

Type
.sql script

Size
217 bytes

Results settings per query
- First 5000 rows (default)
- All rows

Results | Messages

View: Table | Chart | Export results

Search

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3059.991
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
773	Mountain-100 Silver, 44	Mountain Bikes	3399.9900
774	Mountain-100 Silver, 48	Mountain Bikes	3399.9900
775	Mountain-100 Black, 38	Mountain Bikes	3374.9900

00:00:02 Query executed successfully.

Search

ENG IN | 7:27 PM 11/13/2024

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labI...

labs.xtremelabs.io/LabViewerConnection/DetachLabManual?labInst...

Instance ID:6757624
XtremeLabs: DP-203T00-A-CEP [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics

100% Completed

Lab Time Left: 00:51:36

Lab Actions

10. Use Delta Lake with Spark in Azure Synapse Analytics

Delete Azure resources:

- [x] If you've finished exploring Azure Synapse Analytics, you should delete the resources you've created to avoid unnecessary Azure costs.
- [x] 1. Close the Synapse Studio browser tab and return to the Azure portal.
- [x] 2. On the Azure portal, on the **Home** page, select **Resource groups**.
- [x] 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 and storage account for your workspace.
- [x] 4. At the top of the **Overview** page for your resource group, select **Delete resource group**.
- [x] 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.

Previous Exercise

Page: 10/10

Support

XtremeLabs - [DP-203T00-A-M

dp203-qyr3wu5 - Microsoft Azu

portal.azure.com/#@xtremelabs.onmicrosoft.com/resource/subscriptions/9353cc25-ed82-417d-9415-c...

Microsoft Azure

Search resources, services, and docs (G+/)

Copilot

Home > Resource groups >

Deleted resource group dp203-qyr3wu5

Deleted resource group dp203-qyr3wu5

dp203-qyr3wu5

Resource group

Search

- Overview
- Activity log
- Access control (IAM)
- Tags
- Resource visualizer
- Events
- Settings
- Cost Management
- Monitoring
- Automation
- Help

Create | Manage view | Delete resource group | Refresh | Export to CSV

Essentials

JSON View

Resources | Recommendations

Filter for any field... | Type equals **all** | Location equals **all** | Add filter

Showing 0 to 0 of 0 records. Show hidden types

No grouping

List view

Name | Type | Location

No resources match your filters

Try changing or clearing your filters.

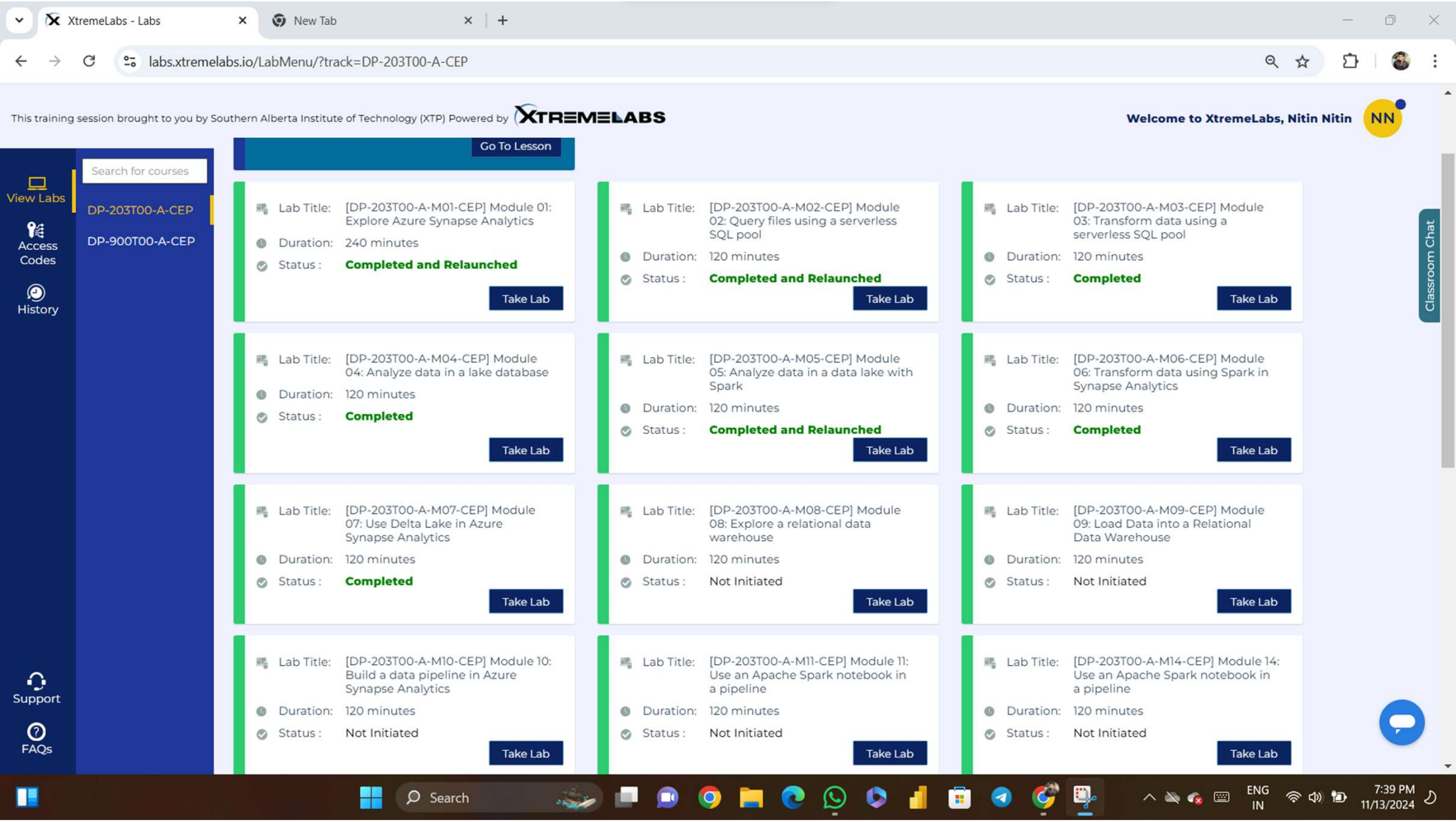
Create resources | Clear filters

Give feedback

Search

ENG
IN

7:37 PM
11/13/2024


XtremeLabs - Labs
New Tab
labs.xtremelabs.io/LabMenu/?track=DP-203T00-A-CEP
This training session brought to you by Southern Alberta Institute of Technology (XTP) Powered by XTREMELABS
Welcome to XtremeLabs, Nitin Nitin
NN
Go To Lesson
View Labs
Access Codes
History
Support
FAQs
Search for courses
DP-203T00-A-CEP
DP-900T00-A-CEP
Classroom Chat
Lab Title: [DP-203T00-A-M01-CEP] Module 01: Explore Azure Synapse Analytics
Duration: 240 minutes
Status : Completed and Relaunched
Take Lab
Lab Title: [DP-203T00-A-M02-CEP] Module 02: Query files using a serverless SQL pool
Duration: 120 minutes
Status : Completed and Relaunched
Take Lab
Lab Title: [DP-203T00-A-M03-CEP] Module 03: Transform data using a serverless SQL pool
Duration: 120 minutes
Status : Completed
Take Lab
Lab Title: [DP-203T00-A-M04-CEP] Module 04: Analyze data in a lake database
Duration: 120 minutes
Status : Completed
Take Lab
Lab Title: [DP-203T00-A-M05-CEP] Module 05: Analyze data in a data lake with Spark
Duration: 120 minutes
Status : Completed and Relaunched
Take Lab
Lab Title: [DP-203T00-A-M06-CEP] Module 06: Transform data using Spark in Synapse Analytics
Duration: 120 minutes
Status : Completed
Take Lab
Lab Title: [DP-203T00-A-M07-CEP] Module 07: Use Delta Lake in Azure Synapse Analytics
Duration: 120 minutes
Status : Completed
Take Lab
Lab Title: [DP-203T00-A-M08-CEP] Module 08: Explore a relational data warehouse
Duration: 120 minutes
Status : Not Initiated
Take Lab
Lab Title: [DP-203T00-A-M09-CEP] Module 09: Load Data into a Relational Data Warehouse
Duration: 120 minutes
Status : Not Initiated
Take Lab
Lab Title: [DP-203T00-A-M10-CEP] Module 10: Build a data pipeline in Azure Synapse Analytics
Duration: 120 minutes
Status : Not Initiated
Take Lab
Lab Title: [DP-203T00-A-M11-CEP] Module 11: Use an Apache Spark notebook in a pipeline
Duration: 120 minutes
Status : Not Initiated
Take Lab
Lab Title: [DP-203T00-A-M14-CEP] Module 14: Use an Apache Spark notebook in a pipeline
Duration: 120 minutes
Status : Not Initiated
Take Lab
Search
ENG IN
7:39 PM 11/13/2024