

Validate connectivity from Spark

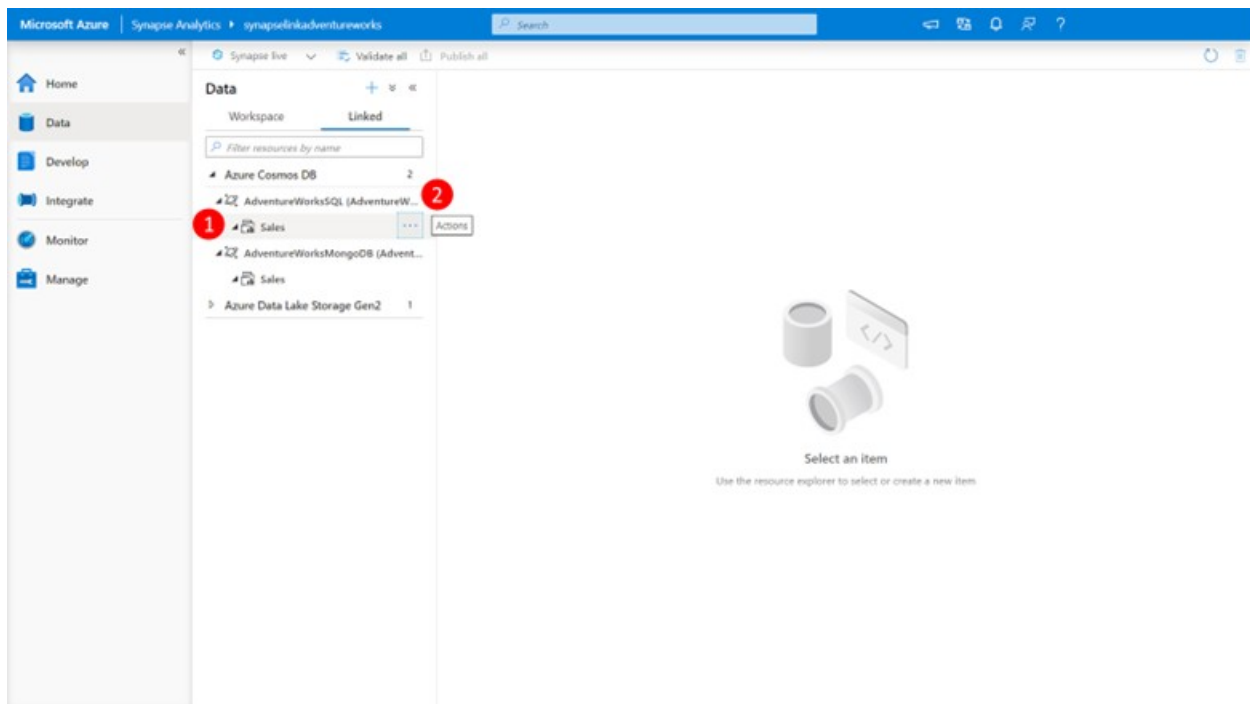
In this reading you can see the steps involved in the process of validating connectivity from Spark.

Note

You are not required to complete the processes, tasks, activities, or steps presented in this example. Your system set-up may differ from the system set-up in the demonstration in this reading. The various samples provided are for illustrative purposes only and it's likely that if you try this out you will encounter issues in your system.

Spark Queries for Cosmos DB Core (SQL) API

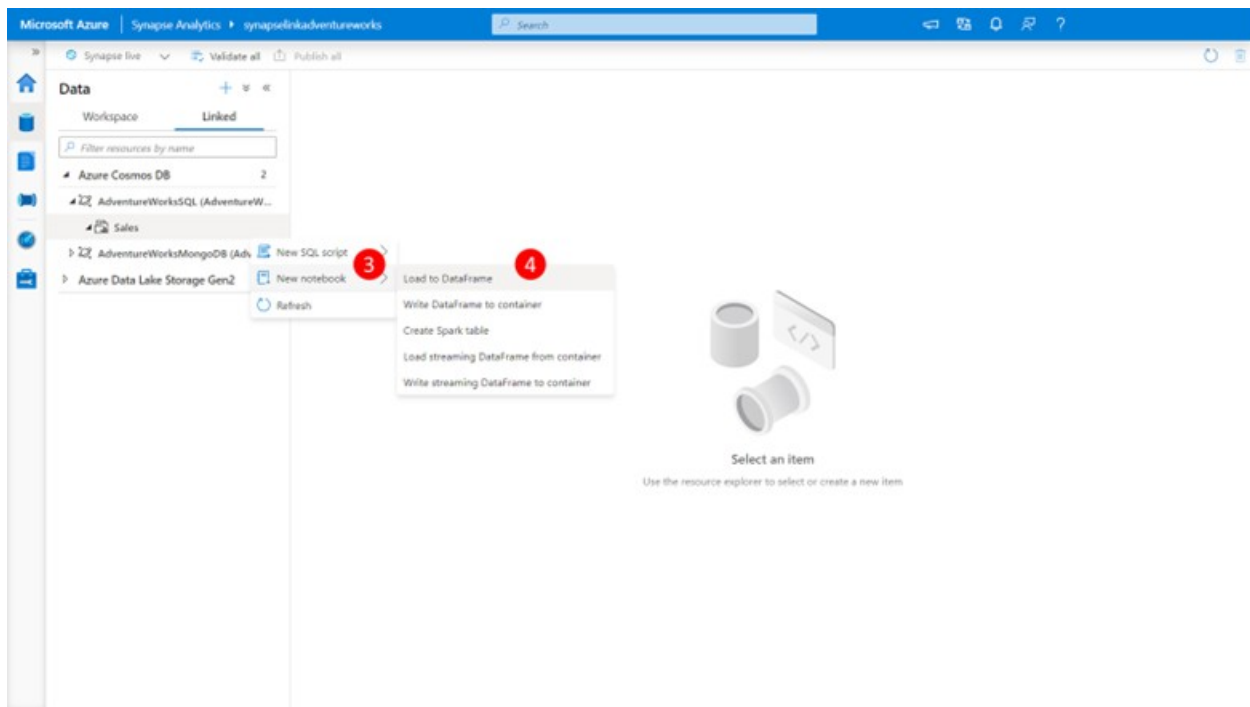
Let's now connect to our Cosmos DB Cosmos DB Core (SQL) API analytical store using Spark and retrieve some data by performing the following steps:



Retrieving data from a Cosmos DB analytical store.

1. Expand the **AdventureWorksSQL linked service** in the explorer view and click on the **Sale container (1)**

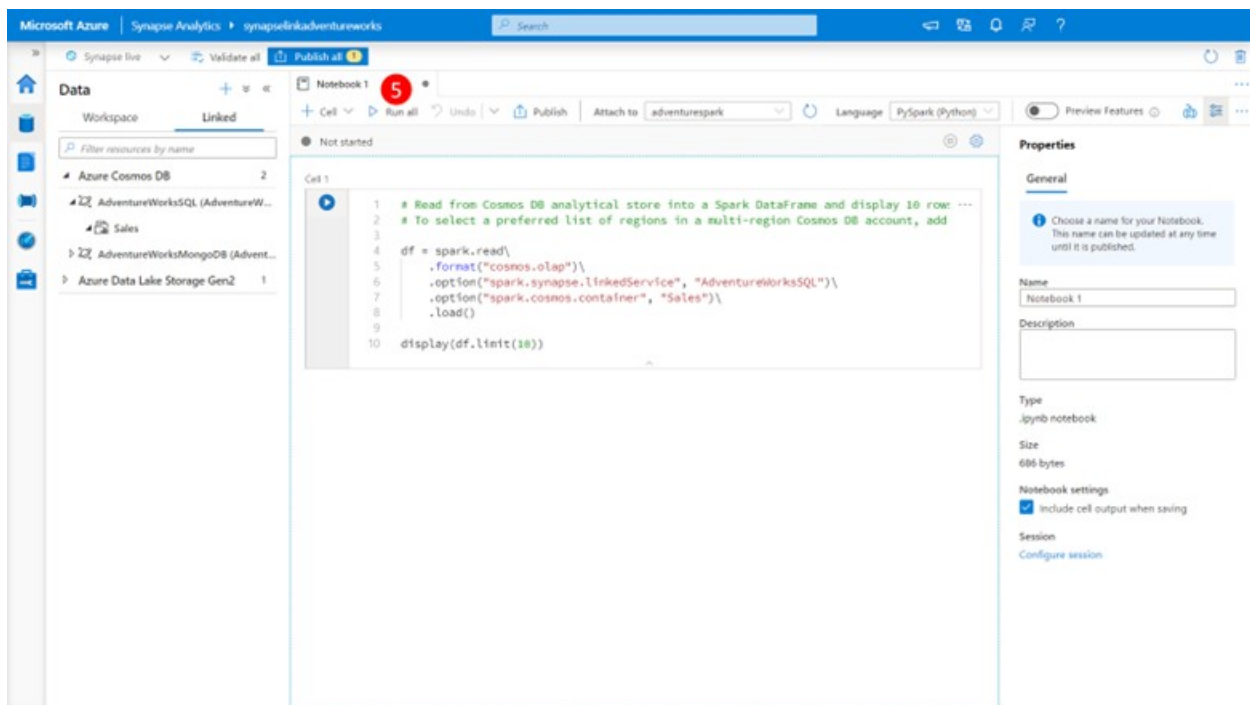
2. Click on the **Actions ellipsis “...”**



Loading data into a DataFrame.

3. Click on **new notebook** to expose the list of notebooks actions.

4. Click on **Load to DataFrame**, to load a prepopulated notebook with a spark query to retrieve the top 10 records from the Linked Server and its associated analytical store.



Running a query to view the top 10 records in an analytical store.

5. Click the **Run All** button on the ribbon to execute the notebook.

Microsoft Azure | Synapse Analytics | synapselinkadventureworks

Workspace | Linked

Filter resources by name

- Azure Cosmos DB 2
- AdventureWorksSQL (AdventureW...
- Sales
- AdventureWorksMongoDB (Advent...
- Azure Data Lake Storage Gen2 1

Notebook 1

Ready

Cell 1

```

1 # Read from Cosmos DB analytical store into a Spark DataFrame and display 10 rows from the DataFrame
2 # To select a preferred list of regions in a multi-region Cosmos DB account, add .option("spark.cosmos.preferredRegion
3
4 df = spark.read\
5     .format("cosmos.olap")\
6     .option("spark.synapse.linkedService", "AdventureworksSQL")\
7     .option("spark.cosmos.container", "Sales")\
8     .load()
9
10 display(df.limit(10))

```

Command executed in 2mins 4bs 365ms by gateway on 11-25-2020 0057:39.211 -08:00

Job execution Succeeded Spark 2 executors 16 cores

View in monitoring Open Spark UI

View Table Chart

_rid	_ts	id	type	name	customerId
mP0TAJ5j9kDA...	1606292304	000C23D8-888C-432E-9213-6473DFA28C5	salesOrder		54A887A7-8DB9-4FAE-A668-AA9F43E26628
mP0TAJ5j9kBA...	1606291881	54A887A7-8DB9-4FAE-A668-AA9F43E26628	customer	Franklin Ye	54A887A7-8DB9-4FAE-A668-AA9F43E26628

Viewing the top 10 records in an analytical store.

You should almost immediately see the query begin to execute and then shortly thereafter receive back a result set **(6)**

Note That for records where data was not defined, such as the name column for the salesOrder record we get back a null value.

Microsoft Azure | Synapse Analytics | synapselinkadventureworks

Workspace | Linked

Filter resources by name

- Azure Cosmos DB 2
- AdventureWorksSQL (AdventureW...
- Sales
- AdventureWorksMongoDB (Advent...
- Azure Data Lake Storage Gen2 1

Notebook 1

Ready

Cell 1

```

1 # Read from Cosmos DB analytical store into a Spark DataFrame and display 10 rows from the DataFrame
2 # To select a preferred list of regions in a multi-region Cosmos DB account, add .option("spark.cosmos.preferredRegion
3
4 df = spark.read\
5     .format("cosmos.olap")\
6     .option("spark.synapse.linkedService", "AdventureworksSQL")\
7     .option("spark.cosmos.container", "Sales")\
8     .load()
9
10 display(df.limit(10))

```

Command executed in 2mins 4bs 365ms by gateway on 11-25-2020 0057:39.211 -08:00

Job execution Succeeded Spark 2 executors 16 cores

View in monitoring Open Spark UI

View Table Chart

me	customerId	address	_etag	orderDate	shipDate	details
54A887A7-8DB9-4FAE-A668-AA9F43E26628			"42001913-0000-...	2014-02-16T00:0...	2014-02-23T00:0...	{ "sku": "BK-R64V", "sku": "BK-R64", "name": "Road", "price": "1120.45", "quantity": "1" }
Franklin Ye	54A887A7-8DB9-4FAE-A668-AA9F43E26628	{ "streetNo": "15850", "streetName": "NE 40th St.", "postcode": "98052" }	"42006809-0000-...			

Viewing JSON results in an analytical store.

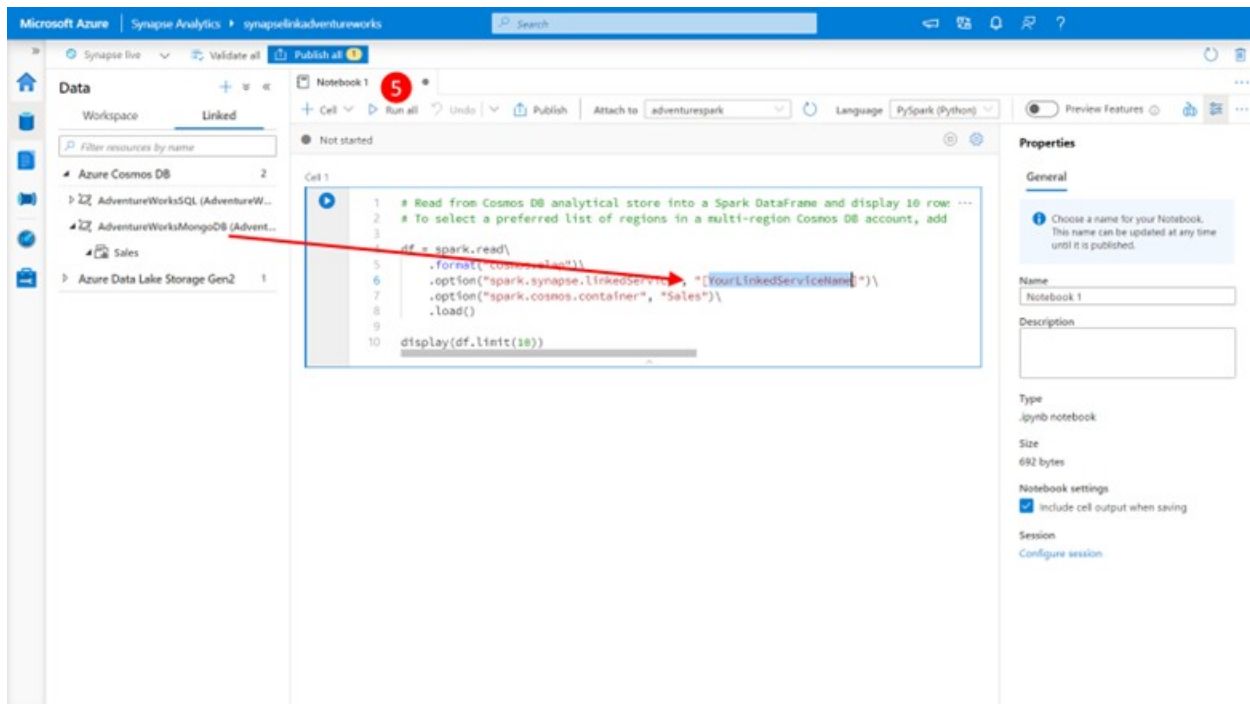
6. Scroll right though the result set

You will see that columns that contain JSON objects, such as address **(8)** and JSON arrays such as detail **(9)** has the JSON as their column content.

Spark queries for Azure Cosmos DB API for MongoDB

Let's now connect to our Azure Cosmos DB API for MongoDB analytical store using Spark and retrieve some data by performing the following steps:

1. Expand the **AdventureWorksMongoDB linked service** in the explorer view and click on the **Sales container**.
2. Click on the **Actions ellipsis "..."**
3. Click on **new notebook** to expose the list of notebooks actions.
4. Click on **Load to DataFrame**, to load a prepopulated notebook with a spark query to retrieve the top 10 records from the Linked Server and its associated analytical store.



Defining a linked service in a Spark query.

5. Update the linked service name and click the **Run All** button on the ribbon to execute the notebook.

Microsoft Azure | Synapse Analytics | synapselinkadventureworks

Workspace | Linked

Filter resources by name

- Azure Cosmos DB 2
 - AdventureWorksSQL (AdventureW...
 - AdventureWorksMongoDB (Advent...
- Sales
- Azure Data Lake Storage Gen2 1

Notebook 1

Cell 1

```

1 # Read from Cosmos DB analytical store into a Spark DataFrame and display 10 rows from the DataFrame
2 # To select a preferred list of regions in a multi-region Cosmos DB account, add .option("spark.cosmos.preferredRegion
3
4 df = spark.read\
5     .format("cosmos.olap")\
6     .option("spark.synapse.linkedService", "AdventureWorksMongoDB")\
7     .option("spark.cosmos.container", "Sales")\
8     .load()
9
10 display(df.limit(10))

```

Command executed in 1min 616ms by gayhope on 11-25-2020 01:15:29:763 -0800

Job execution Succeeded Spark 2 executors 16 cores

View in monitoring Open Spark UI

View Table Chart

_id	_ts	id	_etag	_id	type	name	customerId	address
1f91AP5Vy48DA...	1606293078	NTRBQjg3G7mQ...	"06009356-0000-...	"[string]:"54A887"	"[string]:"customer"	"[string]:"Franklin"	"[string]:"54A887"	"[object]:"54A887"
1f91AP5Vy48EA...	1606293181	MDAwQztrRdgt...	"06009456-0000-...	"[string]:"000C23E"	"[string]:"salesOrder"		"[string]:"54A887"	

Executing a Spark query.

You should almost immediately see the query begin to execute and then shortly thereafter receive back a result set **(6)**

Note That for records where data was not defined, such as the name column for the salesOrder record we get back a null value, and that the name column now contains a JSON fragment with both the data type and value since the MongoDB API uses full fidelity schema mode by default.

Microsoft Azure | Synapse Analytics | synapselinkadventureworks

Workspace | Linked

Filter resources by name

- Azure Cosmos DB 2
 - AdventureWorksSQL (AdventureW...
 - AdventureWorksMongoDB (Advent...
- Sales
- Azure Data Lake Storage Gen2 1

Notebook 1

Cell 1

```

1 # Read from Cosmos DB analytical store into a Spark DataFrame and display 10 rows from the DataFrame
2 # To select a preferred list of regions in a multi-region Cosmos DB account, add .option("spark.cosmos.preferredRegion
3
4 df = spark.read\
5     .format("cosmos.olap")\
6     .option("spark.synapse.linkedService", "AdventureWorksMongoDB")\
7     .option("spark.cosmos.container", "Sales")\
8     .load()
9
10 display(df.limit(10))

```

Command executed in 1min 616ms by gayhope on 11-25-2020 01:15:29:763 -0800

Job execution Succeeded Spark 2 executors 16 cores

View in monitoring Open Spark UI

View Table Chart

_id	type	name	customerId	address	orderDate	shipDate	details
156-0000-...	"[string]:"54A887"	"[string]:"customer"	"[string]:"Franklin"	"[object]:"street"			
156-0000-...	"[string]:"000C23E"	"[string]:"salesOrder"	"[string]:"54A887"		"[string]:"2014-02"	"[string]:"2014-02"	"[array]:"[object]"

Viewing Spark query results.

6. Scroll right though the result set

You will see that columns that contain JSON objects, such as address **(8)** and JSON arrays such as detail **(9)** has the JSON as their column content as well, however this too is expanded to include the data type information.