Azure Synapse Link for Dataverse

Getting Started with Dynamics 365 Sales & Azure Synapse Analytics

Andy Cutler

Original version: February 2023

Updated version: July 2023

Serverlesssql.com

Contents

Configuring Synapse Link for Dataverse Introduction

Part of the Azure Synapse Analytics suite of services is **Synapse Link**. This enables synchronization of data from sources including Cosmos DB, SQL Server, Azure SQL Database, and the Dataverse into a Synapse Analytics workspace and makes it available for querying. This is very useful to consolidate several data sources into a Synapse workspace without needing to setup custom data loading processes.

For example, Synapse Link for SQL enables synchronization of data from Azure SQL Database and SQL Server into Dedicated SQL Pools. Now with **Synapse Link for Dataverse** we can setup data synchronization from Power Apps (including Dynamics 365) into a Synapse Analytics workspace and the data will be available for querying using Serverless SQL Pools and Spark.

In this section we'll be looking at setting up a Synapse Link for Dataverse and **synchronizing Dynamics 365 Sales** data with a Synapse Analytics workspace. We'll also be looking at what is created when the initial setup is run and look at what "near real-time" means, plus any issues we encounter.

Azure Synapse Link Options

There are currently 2 options when configuring Synapse Link. Please note we'll be configuring Synapse Link with Synapse Analytics in this eBook.

- Sync with Synapse Analytics (which includes syncing to an Azure Data Lake Gen2 account)
- Sync with Azure Data Lake Gen2 account

Sync with Synapse Analytics

This option will synchronise Dataverse data to an Azure Data Lake Gen2 storage account and deploy a Lake Database in Synapse Analytics. This Lake Database will hold the Dynamics tables that have been configured for export. There are 2 tables for each entity, a "near-real-time" table, and a "per hour" snapshot table. The "per hour" snapshot table is to minimise any locking during reading (this may occur on the "near-real-time" tables as data is written to CSV files).

What's useful about the Synapse Analytics sync is that each table synchronised from Dynamics is created in the Lake Database with all column names and data types specified. We'll look at the metadata associated with these tables later in this eBook.

Sync with Data Lake

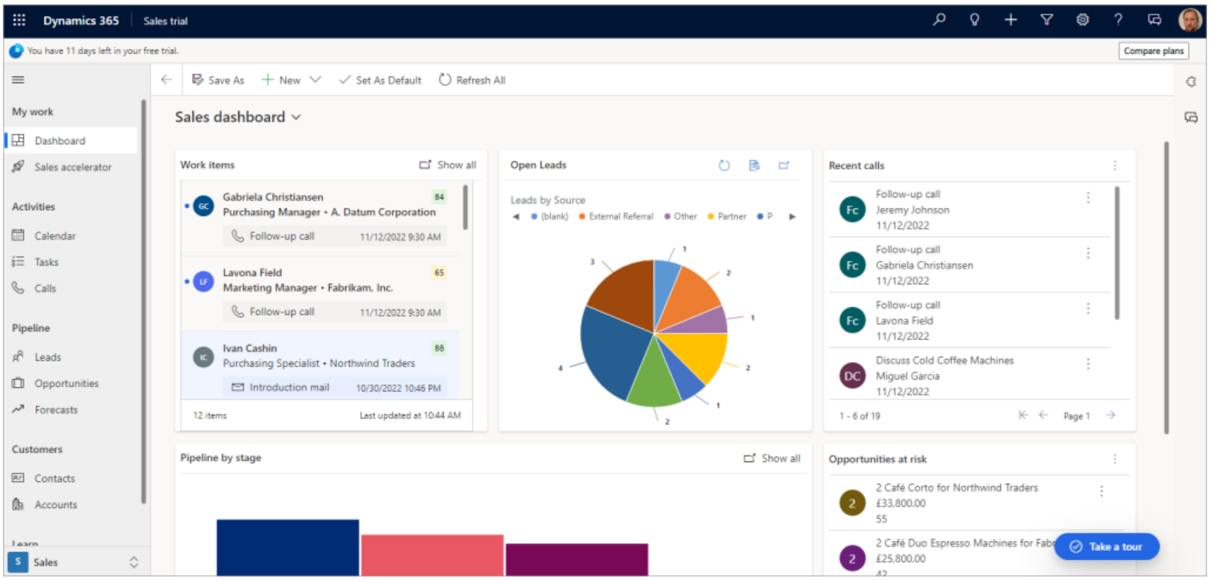
This option will synchronize Dataverse data to an Azure Data Lake Gen2 storage account, this data can then be queried or loading using any service that can connect to an Azure Data Lake Gen2 account and read/process CSV files.

Considerations

It's worth noting that currently the only supported file format when exporting is CSV, this is set to change however with support for Parquet format coming soon (Q4 2022/Q1 2023). It's also worth noting that currently Synapse Link cannot be setup with private endpoints, this is again coming soon (Q4 2022/Q1 2023).

Dynamics 365 Sales

For this walkthrough, I'll be using **Dynamics 365 Sales** as this is a model-driven app and uses the Dataverse. I created a Dynamics 365 Sales trial (30 days) from here. We'll be using the **Contacts** area to add new data and amend existing data, we'll then see this data synchronised in Synapse Analytics.



You have 11 days left in your free trial.

Gabriela Christiansen 84
Purchasing Manager • A. Datum Corporation
Follow-up call 11/12/2022 9:30 AM

Lavona Field 65
Marketing Manager • Fabrikam, Inc.
Follow-up call 11/12/2022 9:30 AM

Ivan Cashin 88
Purchasing Specialist • Northwind Traders
Introduction mail 10/30/2022 10:46 PM

12 items Last updated at 10:44 AM

Follow-up call
Jeremy Johnson
11/12/2022

Follow-up call
Gabriela Christiansen
11/12/2022

Follow-up call
Lavona Field
11/12/2022

Discuss Cold Coffee Machines
Miguel Garcia
11/12/2022

1 - 6 of 19 Page 1

2 Café Corto for Northwind Traders
£33,800.00
55

2 Café Duo Espresso Machines for Fabr
£25,800.00
42

Example Sales dashboard in Dynamics 365 Sales.

Walkthrough: Configuring Synapse Link with a Synapse Workspace

In this section we'll walkthrough how to setup a Synapse Link for Dataverse with a Synapse Analytics workspace. This includes the licensing and permissions required to perform the setup.

Security Requirements & Pre-requisites

An **Azure Synapse Analytics workspace** is required for this walkthrough, please refer to this blog in how to setup a new Synapse workspace.

An Azure storage account setup as a Data Lake Gen2 account needs to be created and added as a linked service to the Synapse workspace.

A **Dynamics 365 Sales** environment is also required, for this tutorial I created a 30-day Sales trial in my own tenant. The Dynamics environment must be in the same region as the Synapse workspace and the Azure Data Lake Gen2 account. E.G for this tutorial the Dynamics/Power Apps environment is in UK West and the Synapse/Storage account is in UK South region.

The user who logs into Power Apps to perform the configuration needs to be licensed via the Office 365 admin portal, and the user requires a license to access Dynamics 365.

Licensing

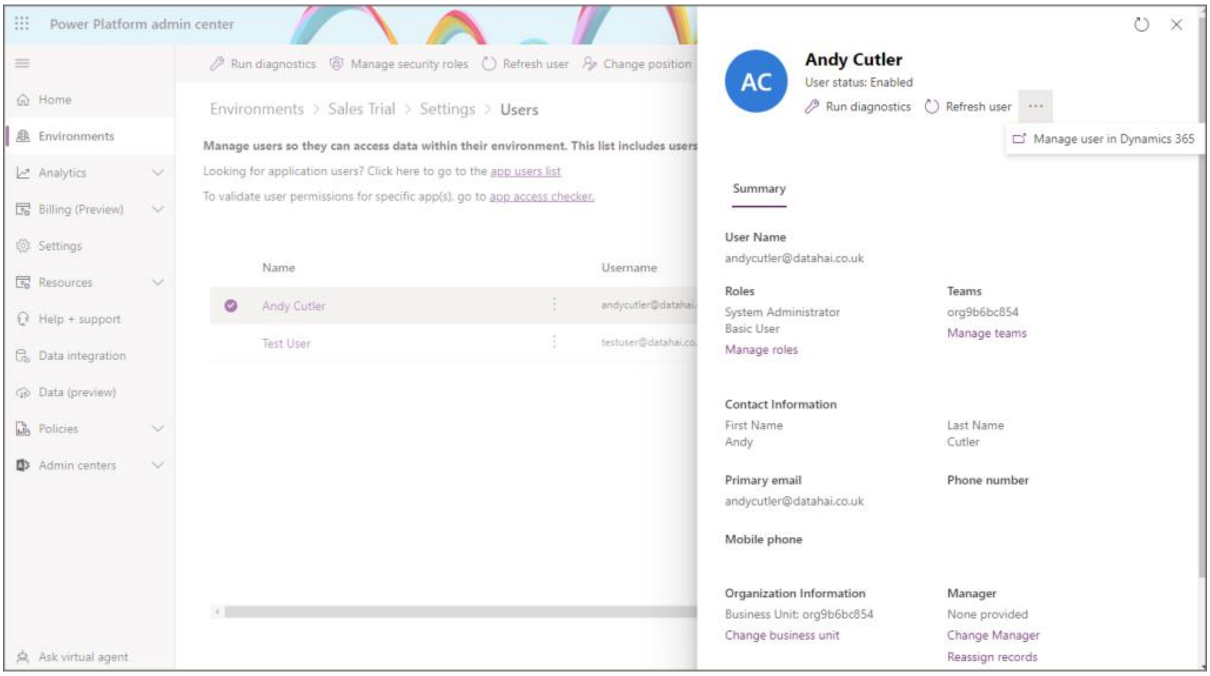
- Power Apps Developer (free) – this can be assigned to the user from the Office 365 admin area.
- Dynamics 365 – a license was automatically granted when I created a trial.

Permissions

The following permissions across Dynamics 365 and Azure are required for the user setting up the Synapse Link with Synapse Analytics.

Dynamics 365

For this tutorial, the user was allocated to the **System Administrator** role in Dynamics 365. You can see user permissions for Dynamics within the Power Platform admin centre in **Environments** > **Your Environment** > **Settings** > **Users + Permissions** > **Users**. Then if you select a user, you can click the **Manage user in Dynamics 365**. Once in Dynamics, click **Manage Roles** on the top menu bar.



Manage users so they can access data within their environment. This list includes users

Looking for application users? Click here to go to the app users list

To validate user permissions for specific app(s), go to app access checker.

Name	Username
Andy Cutler	andycutler@datahai.
Test User	testuser@datahai.co

Andy Cutler

User status: Enabled

User Name
andycutler@datahai.co.uk

Roles
System Administrator
Basic User

Teams
org9b6bc854

Contact Information

First Name
Andy

Last Name
Cutler

Primary email
andycutler@datahai.co.uk

Phone number

Mobile phone

Organization Information
Business Unit: org9b6bc854

Manager
None provided

Azure & Synapse Analytics

- **Resource Group**: Reader
- **Storage Account**:
  - Owner or Role Based Access Control Administrator (Preview)
  - Storage Blob Data Contributor.
- **Synapse Workspace**: Workspace administrator

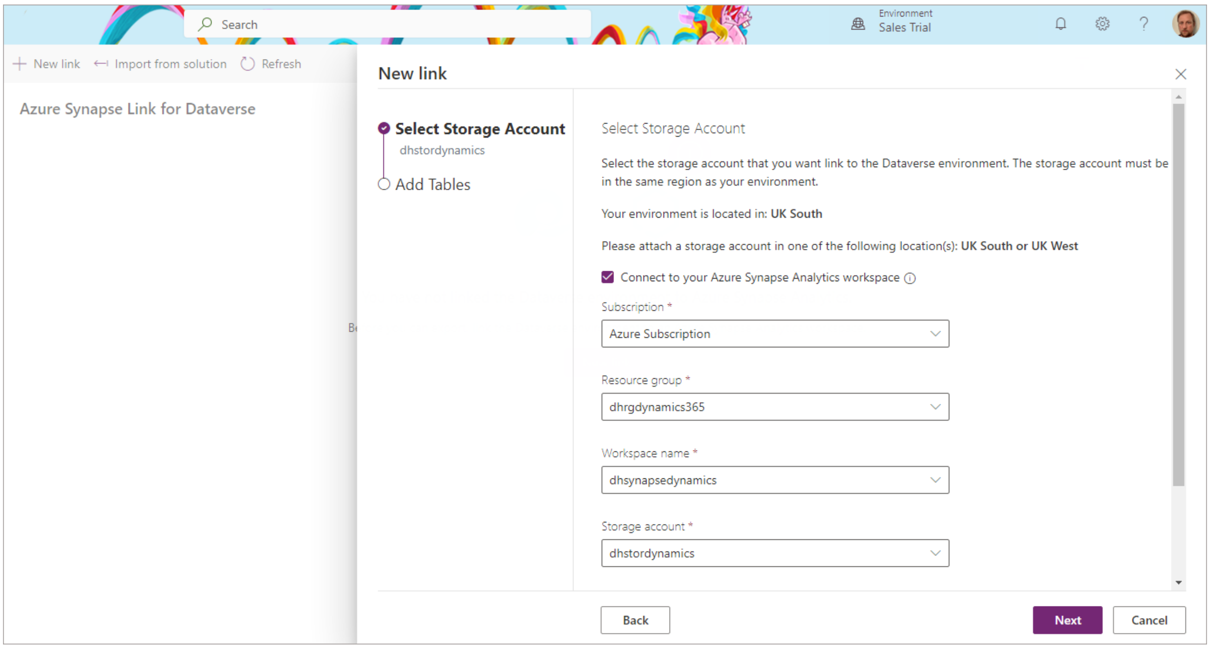
Owner is specified in the documentation here, but I have found that giving the user the **Role Based Access Control Administrator (Preview)** is enough and provides the lowest level if security. However, as this role is in preview, you may not wish to use it. The user can be removed from the Resource Group/Storage Account/Synapse Workspace after setup as it's no longer required.

Setup Synapse Link

- Login to Power Apps
- Select the relevant **environment** from the top-right menu
- On the left menu, select **Dataverse** -> **Azure Synapse Link**
- Click **New Link** and enter the following information:
  - Enable **Connect to your Azure Synapse Analytics workspace**
  - **Subscription**: Select the appropriate subscription
  - **Resource group**: Select the resource group the Synapse workspace is in
  - **Synapse workspace**: Select the specific Synapse workspace
  - **Storage account**: Select the appropriate storage account to use
- Click **Next**

**** Update April 2023 ****

The location of the Synapse Link item may not be immediately available on the left side menu. Hover over **Discover** then click the **Discover all** button. You should then see **Synapse Link** under **Data Management**, you can then pin the item to the menu.



New link

- Select Storage Account
  dhstordynamics
- Add Tables

Select Storage Account

Select the storage account that you want link to the Dataverse environment. The storage account must be in the same region as your environment.

Your environment is located in: **UK South**

Please attach a storage account in one of the following location(s): **UK South or UK West**

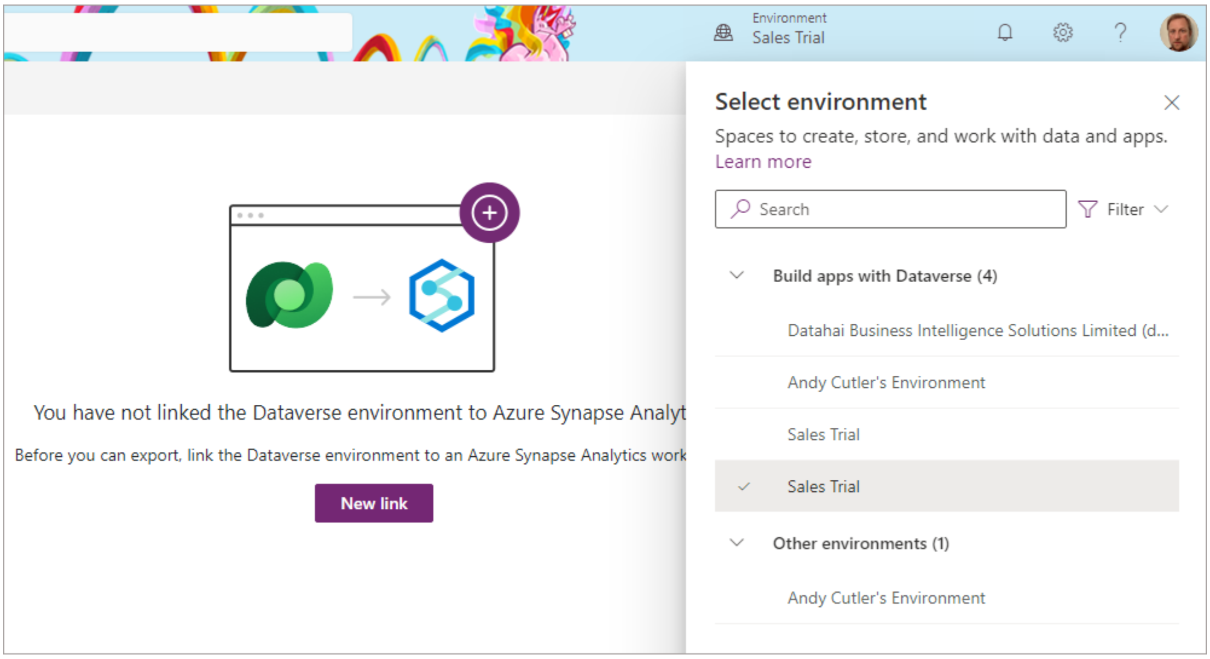
☑ Connect to your Azure Synapse Analytics workspace

Subscription *
Azure Subscription

Resource group *
dhrgdynamics365

Workspace name *
dhsynapsedynamics

Storage account *
dhstordynamics



You have not linked the Dataverse environment to Azure Synapse Analyt

Before you can export, link the Dataverse environment to an Azure Synapse Analytics work

Select environment

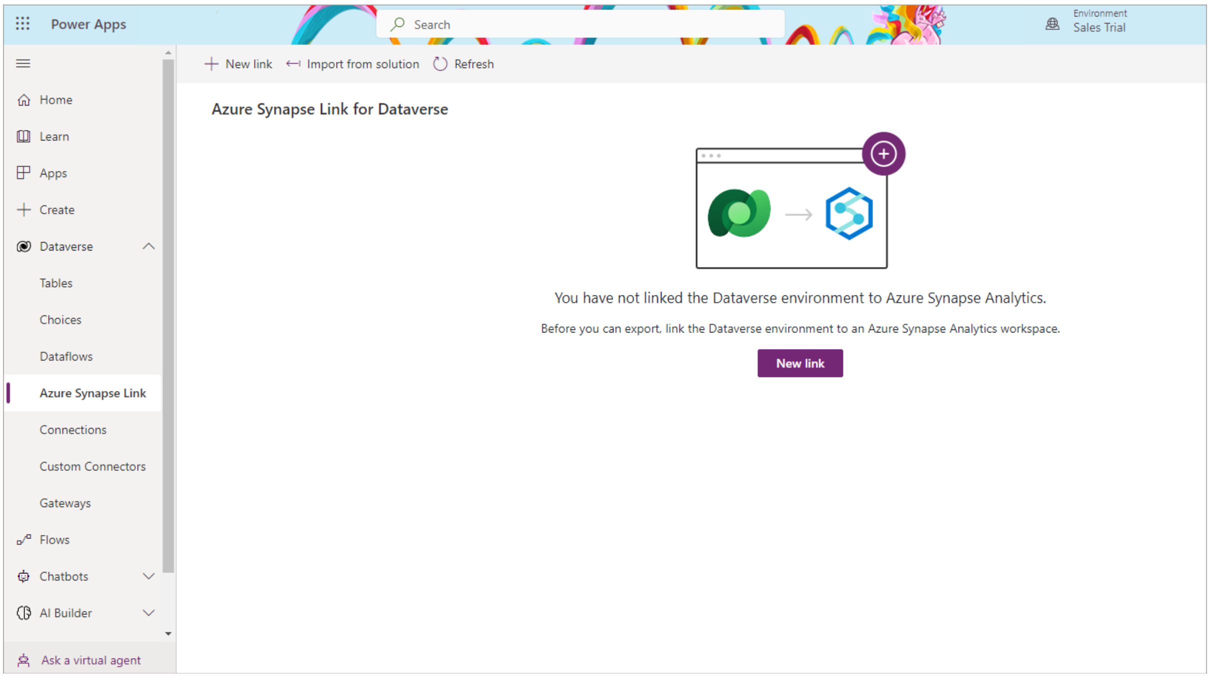
Spaces to create, store, and work with data and apps. Learn more

Build apps with Dataverse (4)

- Datahai Business Intelligence Solutions Limited (d...
- Andy Cutler's Environment
- Sales Trial
- ✓ Sales Trial

Other environments (1)

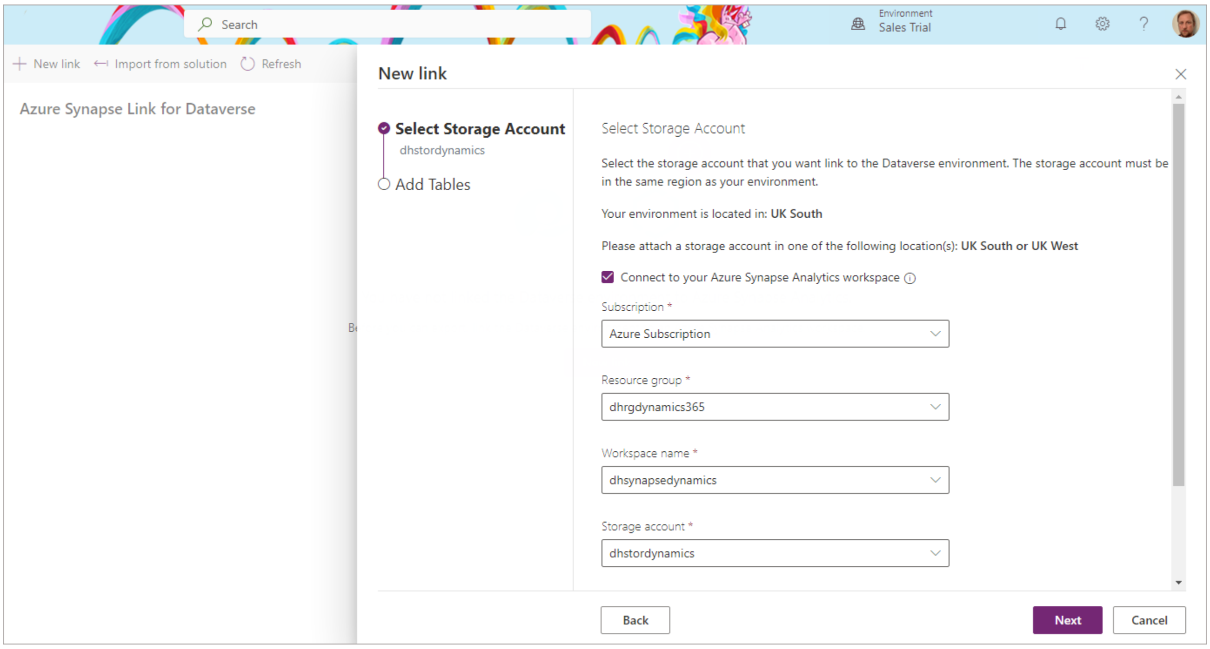
- Andy Cutler's Environment



Azure Synapse Link for Dataverse

You have not linked the Dataverse environment to Azure Synapse Analytics.

Before you can export, link the Dataverse environment to an Azure Synapse Analytics workspace.



New link

- Select Storage Account
  dhstordynamics
- Add Tables

Select Storage Account

Select the storage account that you want link to the Dataverse environment. The storage account must be in the same region as your environment.

Your environment is located in: **UK South**

Please attach a storage account in one of the following location(s): **UK South or UK West**

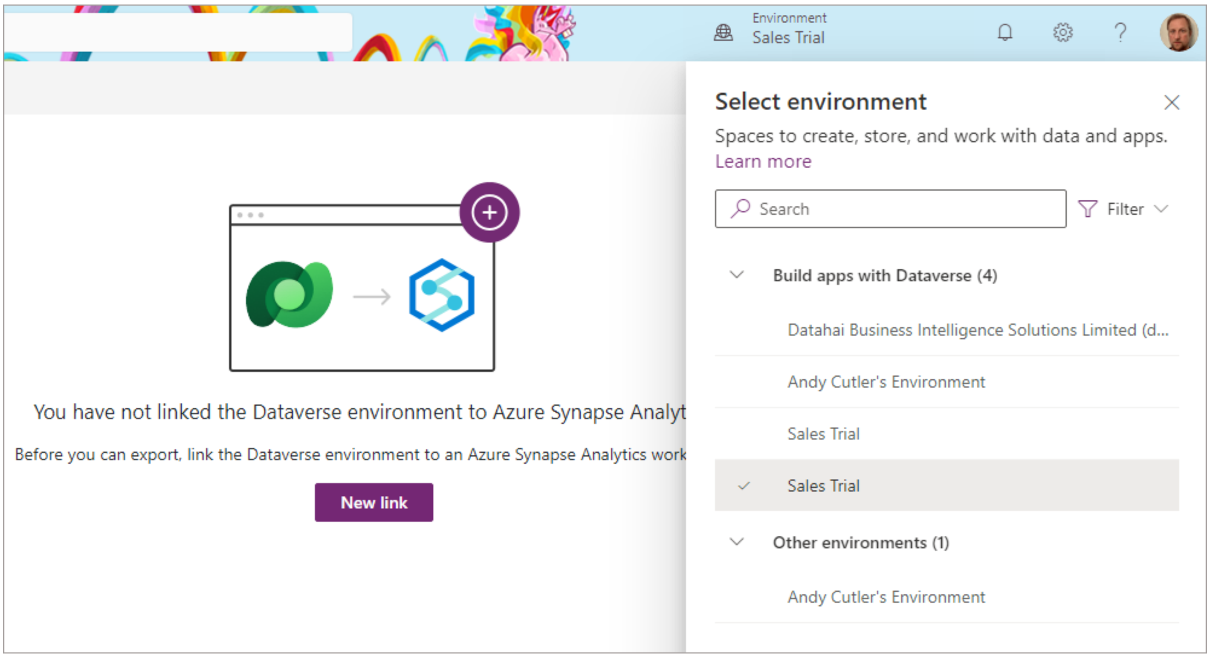
☑ Connect to your Azure Synapse Analytics workspace

Subscription *
Azure Subscription

Resource group *
dhrgdynamics365

Workspace name *
dhsynapsedynamics

Storage account *
dhstordynamics



Environment
Sales Trial

You have not linked the Dataverse environment to Azure Synapse Analyt

Before you can export, link the Dataverse environment to an Azure Synapse Analytics work

New link

Select environment

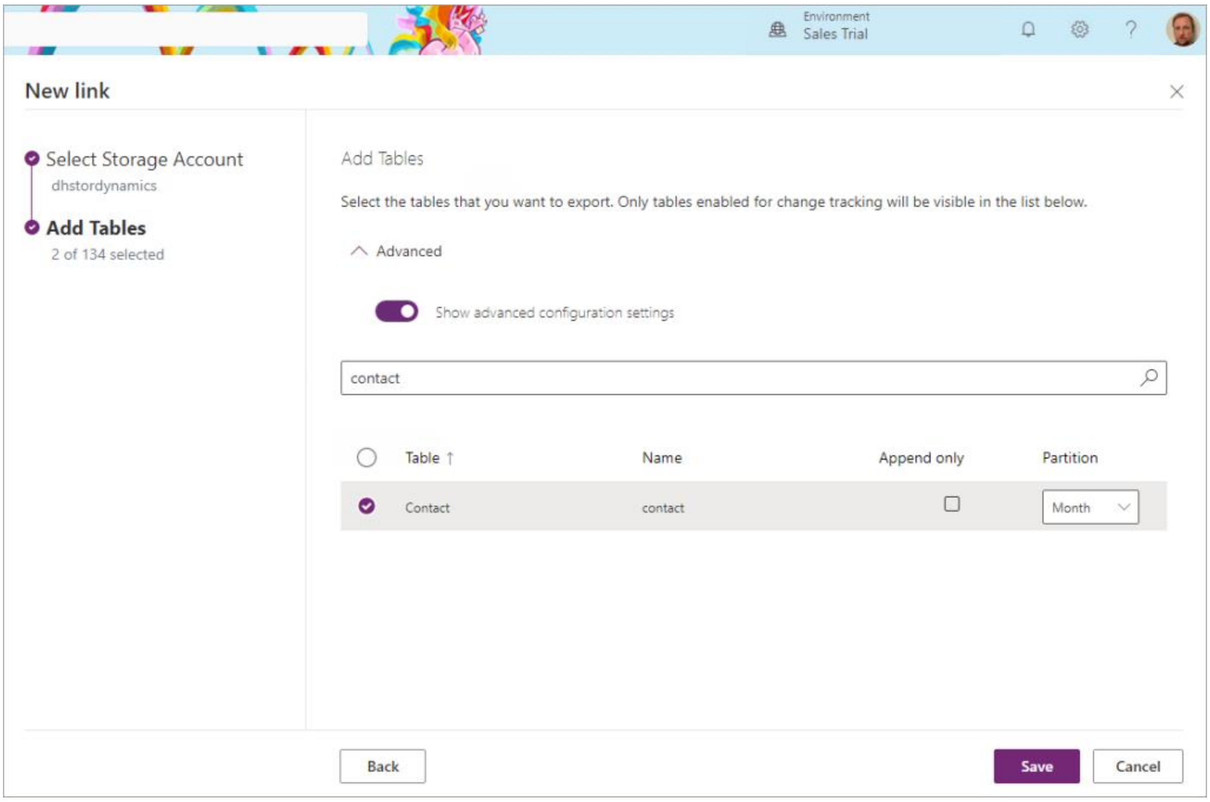
Spaces to create, store, and work with data and apps. Learn more

Search Filter

Build apps with Dataverse (4)
- Datahai Business Intelligence Solutions Limited (d...
- Andy Cutler's Environment
- Sales Trial
- ✓ Sales Trial

Other environments (1)
- Andy Cutler's Environment

- On the **Add Tables** screen, add relevant tables for synchronisation. In this scenario we'll select the **Contact** table.
- Please note we'll look at **Advanced** configuration in another section as this deals with **Append** only setup (no data gets deleted from Synapse/Data Lake when source Dynamics data is deleted).
- Once all relevant tables have been selected, click **Save**.



Environment
Sales Trial

New link

Select Storage Account
dhstordynamics

Add Tables
2 of 134 selected

Add Tables

Select the tables that you want to export. Only tables enabled for change tracking will be visible in the list below.

Advanced

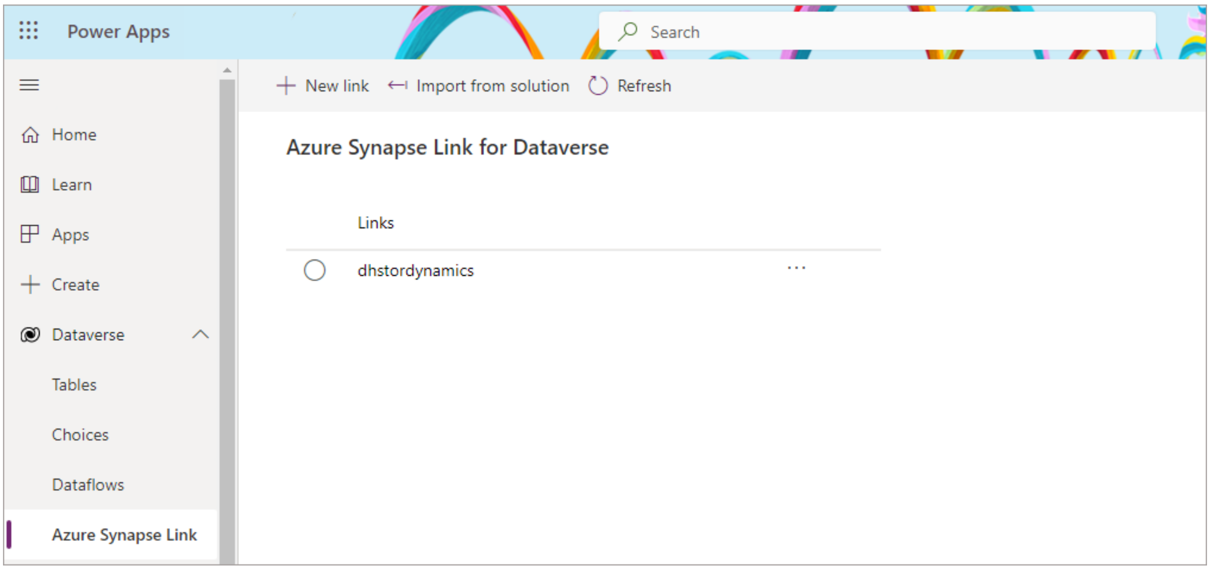
Show advanced configuration settings

contact

	Table ↑	Name	Append only	Partition
●	Contact	contact	□	Month

Back Save Cancel

If the setup has been successful, we'll see the link visible in the Azure Synapse Link area.

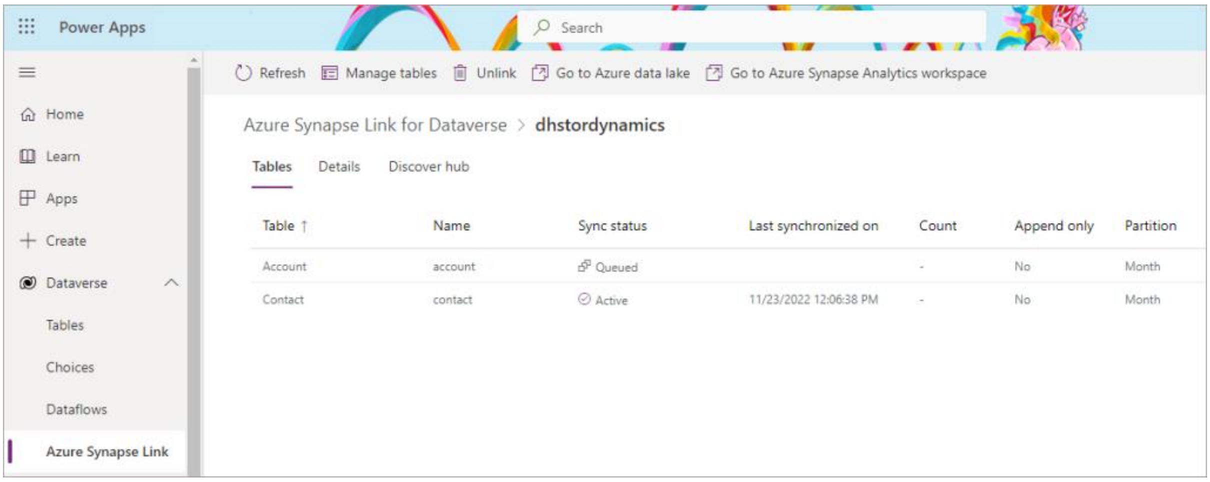


Azure Synapse Link for Dataverse

	Links
○	dhstordynamics

Initial Synchronisation

Once the initial setup has been started, we can click on the Synapse link name and look at the status of the table synchronisation. Now I have seen initial synchronisations take up to 45 minutes if there are a lot of tables, I tested with ~300 tables in one scenario and the total time to sync was around 1 hour. So be warned that the number of tables can affect the initial sync time. It may be better to start with a few tables and then add more tables later (thanks to Scott Sewell at Microsoft for advice).



Azure Synapse Link for Dataverse > dhstordynamics

Table ↑	Name	Sync status	Last synchronized on	Count	Append only	Partition
Account	account	Queued		-	No	Month
Contact	contact	Active	11/23/2022 12:06:38 PM	-	No	Month

Subsequent Refreshes

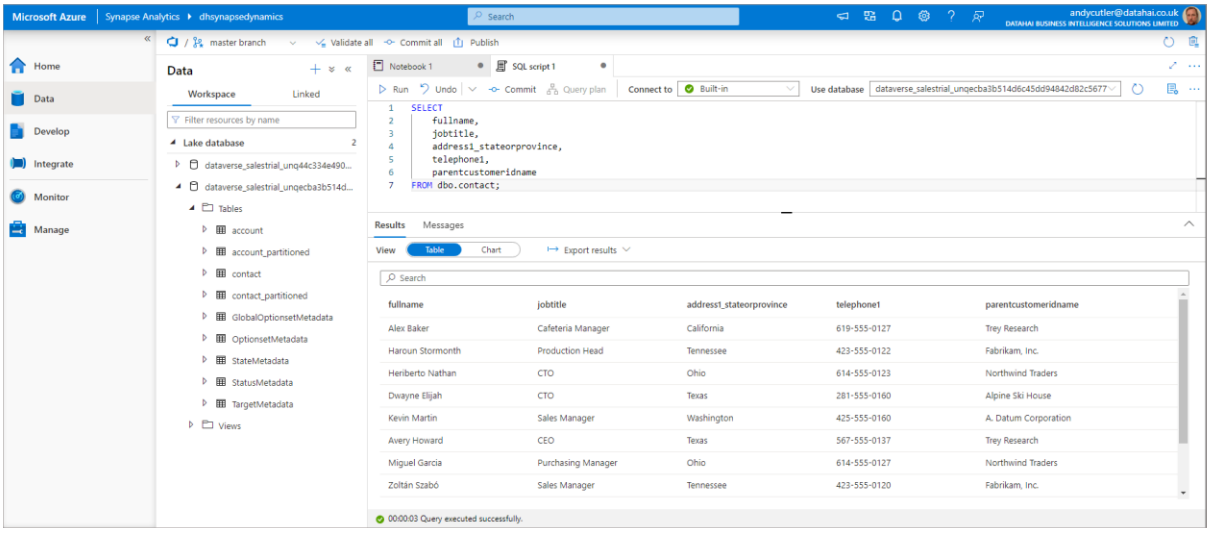
Once the initial synchronisation has completed, the **sync status** should show as **Active.** Any data changes within Dynamics 365 will then be automatically synchronised. The time taken to synchronise data will depend on how much data has changed, this is a push process so changed data gets added to a queue before being synchronised with Synapse/Data Lake. In my testing a small set of data changes (handful of rows in the Contacts table) took around 1 minute. Bulk data changes in Dynamics may make the "near real-time" appear a little slower, this could be up to 15 minutes.

Viewing Data in Synapse Analytics Lake Database

When the Synapse Link is created, it will create a **Lake Database** in Synapse Analytics that we can use to query using **Serverless SQL Pools** (we can also use Spark). To do this:

- Log into the Synapse Analytics workspace (link)
- Click on the **Data** tab on the left menu
- Expand **Lake database** and you should see the lake database that was created during the synapse link setup.
- Expand **Tables** and you should now see the tables selected for export, plus metadata tables (E.G **StateMetadata**)
- We can now click on the **Develop** tab and create a new **SQL script** and start querying the tables.

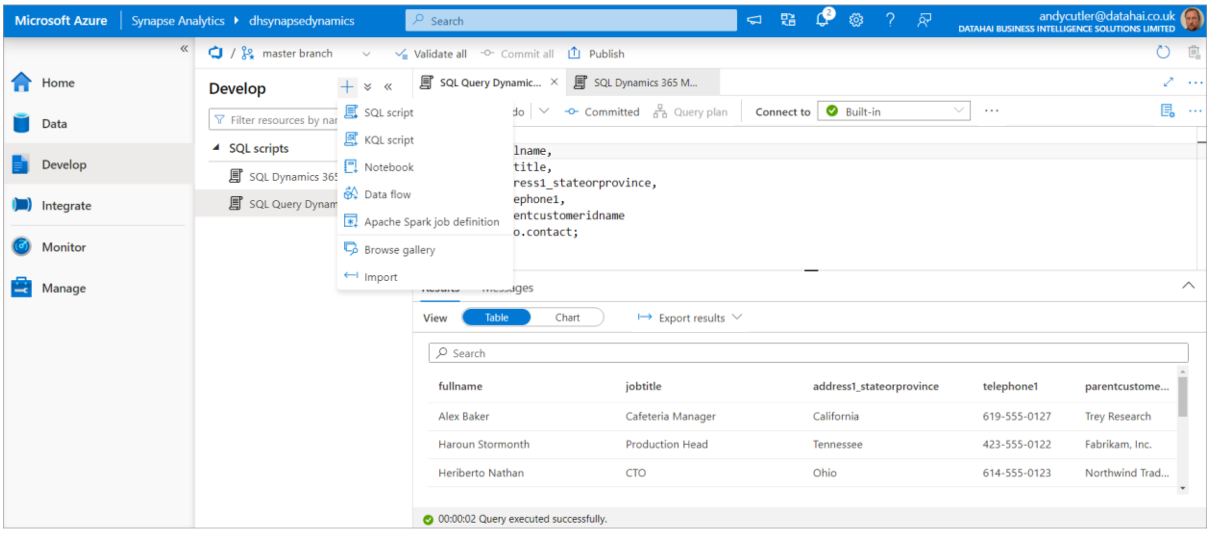
Scroll through images for setup example.



```
SELECT
    fullname,
    jobtitle,
    address1_stateorprovince,
    telephone1,
    parentcustomeridname
FROM dbo.contact;
```

fullname	jobtitle	address1_stateorprovince	telephone1	parentcustomeridname
Alex Baker	Cafeteria Manager	California	619-555-0127	Trey Research
Haroun Stormonth	Production Head	Tennessee	423-555-0122	Fabrikam, Inc.
Heriberto Nathan	CTO	Ohio	614-555-0123	Northwind Traders
Dwayne Elijah	CTO	Texas	281-555-0160	Alpine Ski House
Kevin Martin	Sales Manager	Washington	425-555-0160	A. Datum Corporation
Avery Howard	CEO	Texas	567-555-0137	Trey Research
Miguel Garcia	Purchasing Manager	Ohio	614-555-0127	Northwind Traders
Zoltán Szabó	Sales Manager	Tennessee	423-555-0120	Fabrikam, Inc.

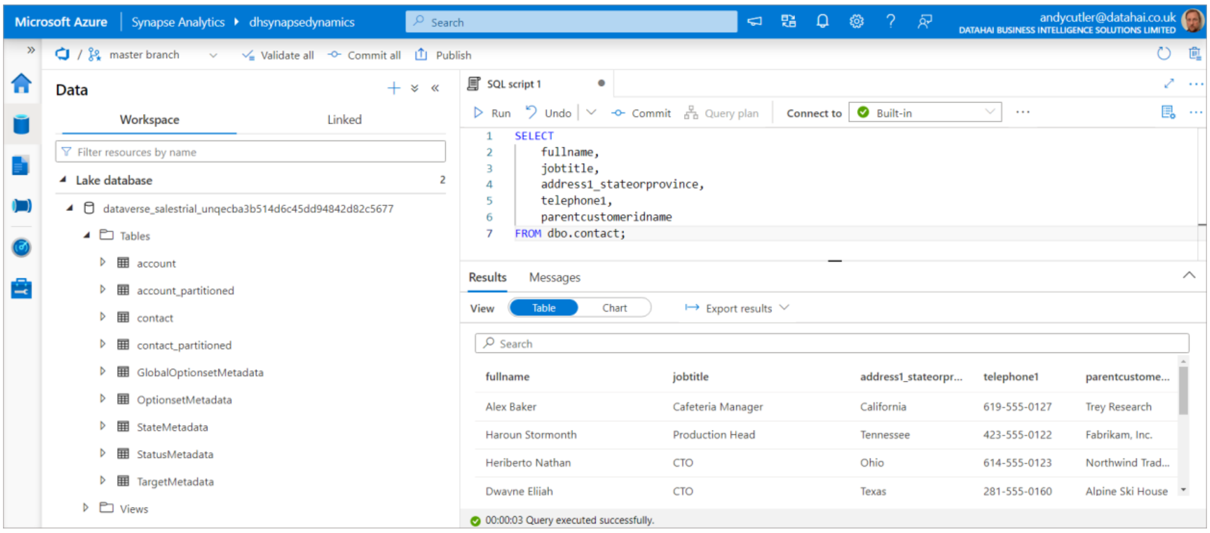
00:00:03 Query executed successfully.



```
lname,
title,
ress1_stateorprovince,
ephone1,
entcustomeridname
o.contact;
```

fullname	jobtitle	address1_stateorprovince	telephone1	parentcustome...
Alex Baker	Cafeteria Manager	California	619-555-0127	Trey Research
Haroun Stormonth	Production Head	Tennessee	423-555-0122	Fabrikam, Inc.
Heriberto Nathan	CTO	Ohio	614-555-0123	Northwind Trad...

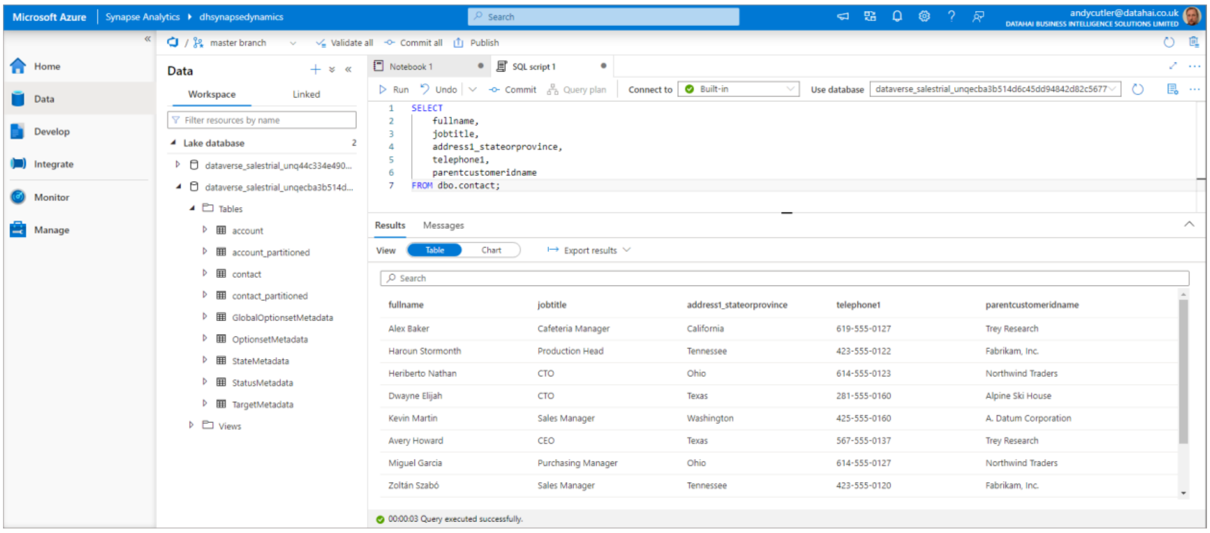
00:00:02 Query executed successfully.



```
SELECT
    fullname,
    jobtitle,
    address1_stateorprovince,
    telephone1,
    parentcustomeridname
FROM dbo.contact;
```

fullname	jobtitle	address1_stateorpr...	telephone1	parentcustome...
Alex Baker	Cafeteria Manager	California	619-555-0127	Trey Research
Haroun Stormonth	Production Head	Tennessee	423-555-0122	Fabrikam, Inc.
Heriberto Nathan	CTO	Ohio	614-555-0123	Northwind Trad...
Dwayne Elijah	CTO	Texas	281-555-0160	Alpine Ski House

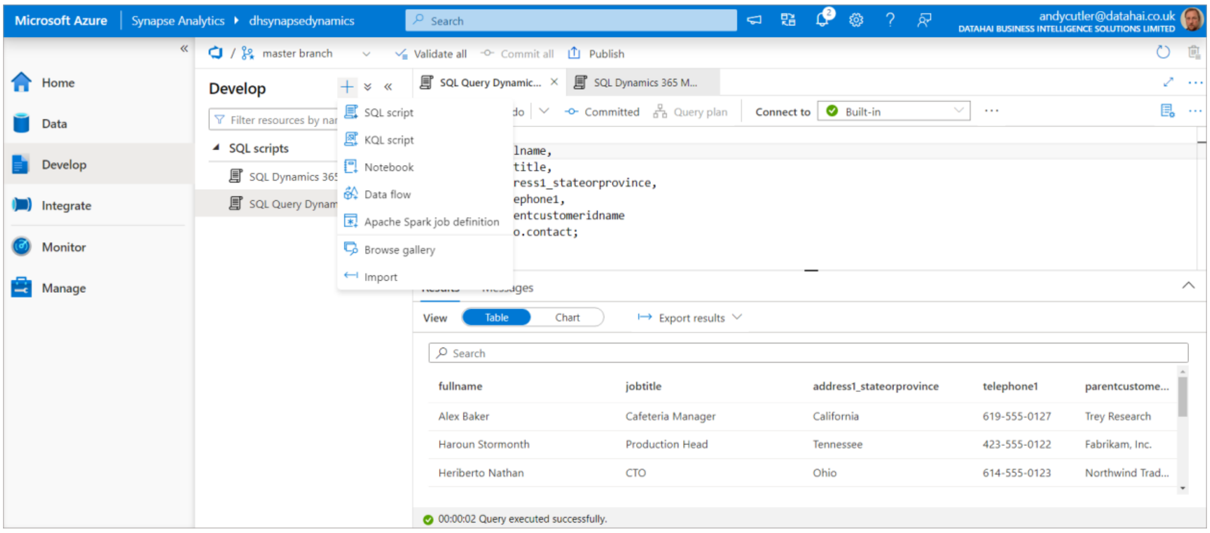
00:00:03 Query executed successfully.



```sql
SELECT
    fullname,
    jobtitle,
    address1_stateorprovince,
    telephone1,
    parentcustomeridname
FROM dbo.contact;
```

fullname	jobtitle	address1_stateorprovince	telephone1	parentcustomeridname
Alex Baker	Cafeteria Manager	California	619-555-0127	Trey Research
Haroun Stormonth	Production Head	Tennessee	423-555-0122	Fabrikam, Inc.
Heriberto Nathan	CTO	Ohio	614-555-0123	Northwind Traders
Dwayne Elijah	CTO	Texas	281-555-0160	Alpine Ski House
Kevin Martin	Sales Manager	Washington	425-555-0160	A. Datum Corporation
Avery Howard	CEO	Texas	567-555-0137	Trey Research
Miguel Garcia	Purchasing Manager	Ohio	614-555-0127	Northwind Traders
Zoltán Szabó	Sales Manager	Tennessee	423-555-0120	Fabrikam, Inc.

00:00:03 Query executed successfully.



```
lname,
title,
ress1_stateorprovince,
ephone1,
entcustomeridname
o.contact;
```

fullname	jobtitle	address1_stateorprovince	telephone1	parentcustome...
Alex Baker	Cafeteria Manager	California	619-555-0127	Trey Research
Haroun Stormonth	Production Head	Tennessee	423-555-0122	Fabrikam, Inc.
Heriberto Nathan	CTO	Ohio	614-555-0123	Northwind Trad...

00:00:02 Query executed successfully.

Table Metadata

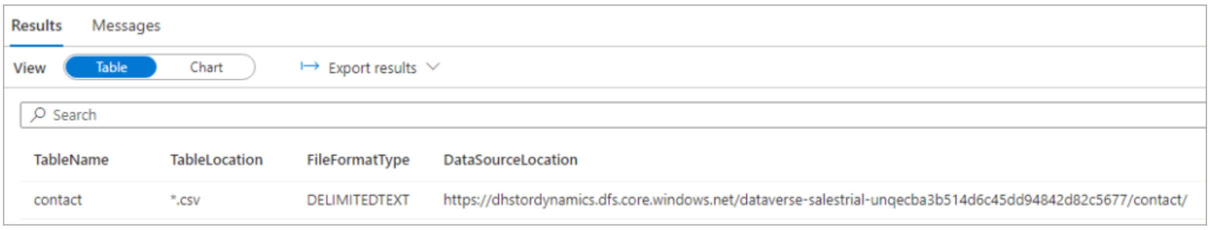
There are 2 tables for each source table in Dynamics, we can see that both the **account** and **contact** tables exist in the Lake database, and we also have **account_partitioned** and **contact_partitioned**. The partitioned tables exist to limit any potential locking issues when reading data from the near real-time tables as the synchronisation window is much larger, usually each hour.

We can look at the metadata associated with each table to understand which location/folder in the data lake it is referencing. NB the _partitioned tables are not visible via the system views (I have an outstanding question about this). The SQL script below is part of my **serverlesssqlpooltools** script library on GitHub here called **externaltablemetadata.sql**

Contact Table Metadata

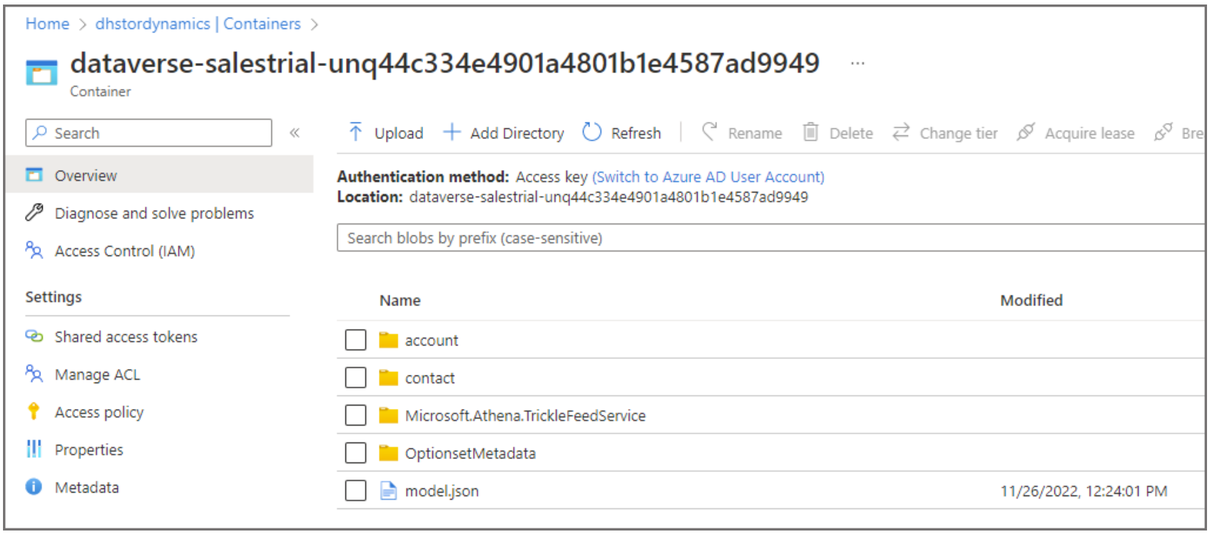
If we run the SQL below on the Lake database, we can use the relevant system views to view the table metadata including its source data format and its location in the data lake.

```
SELECT
            et.[name] AS TableName,
            et.[location] AS TableLocation,
            ef.[format_type] AS FileFormatType,
            es.[location] AS DataSourceLocation
FROM sys.external_tables et
INNER JOIN sys.external_file_formats ef ON ef.file_format_id = et.file_format_id
INNER JOIN sys.external_data_sources es ON es.data_source_id = et.data_source_id
WHERE et.[name] = 'contact'
ORDER BY et.[name]
```



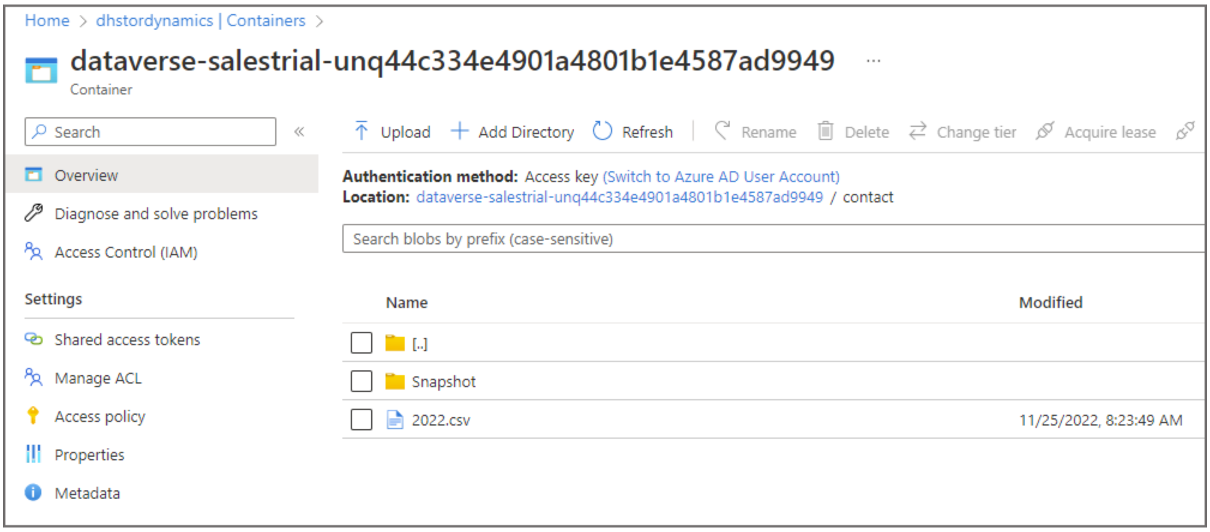
TableName	TableLocation	FileFormatType	DataSourceLocation
contact	*.csv	DELIMITEDTEXT	https://dhstordynamics.dfs.core.windows.net/dataverse-salestrial-unqecba3b514d6c45dd94842d82c5677/contact/

We can see the location of the CSV files in the Data Lake, within the **contact** folder. If we look at the storage account itself, we can see the CSV files within the folder. If we browse to the container within the Azure storage account, we can see folders that represent each table we chose to sync.



Authentication method: Access key (Switch to Azure AD User Account)
Location: dataverse-salestrial-unq44c334e4901a4801b1e4587ad9949

	Name	Modified
□	account	
□	contact	
□	Microsoft.Athena.TrickleFeedService	
□	OptionsetMetadata	
□	model.json	11/26/2022, 12:24:01 PM



dataverse-salestrial-unq44c334e4901a4801b1e4587ad9949

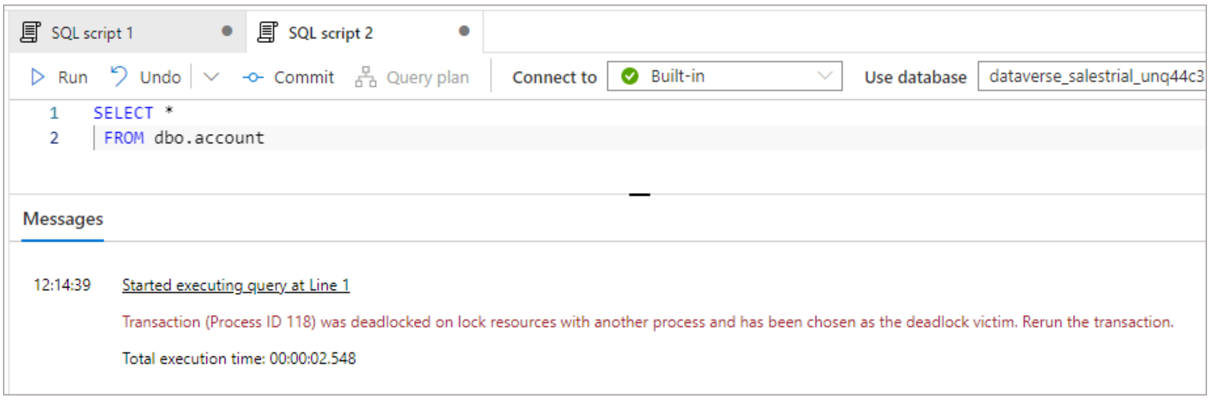
Authentication method: Access key (Switch to Azure AD User Account)
Location: dataverse-salestrial-unq44c334e4901a4801b1e4587ad9949 / contact

	Name	Modified
□	[..]	
□	Snapshot	
□	2022.csv	11/25/2022, 8:23:49 AM

The snapshot folder is ignored by the base table but is used by the _partitioned tables.

Locking

Yes…the dreaded locking. And it happens here too… As the data format in the Data Lake is CSV, when this is being written to by the Synapse Link process it will be locked while writing. There is no way to control this locking behaviour using the Lake Database tables that are created. However, there is the ability to query the CSV data using OPENROWSET which has options to specify reading uncommitted data. This isn't very desirable though and it's best to have some form of retry policy for any services connecting to Synapse to query the data.



Connect to: Built-in
Use database: dataverse_salestrial_unq44c3

```sql
SELECT *
 FROM dbo.account
```

```
12:14:39   Started executing query at Line 1
           Transaction (Process ID 118) was deadlocked on lock resources with another process and has been chosen as the deadlock victim. Rerun the transaction.
           Total execution time: 00:00:02.548
```

What is “Near” Real-time?

The documentation talks about **near real-time** but what can we expect in terms on source data in Dynamics being available for querying in Synapse? The answer….it depends… Synapse Link is a push-based process which adds changed data to a queue. How fast this appears in Synapse depends on how much data is added to the queue. In my testing a handful of record changes such as adding a new Contact and amending an Account sees the data being available in Synapse in under 1 minute. Bulkier updates may see this take longer.

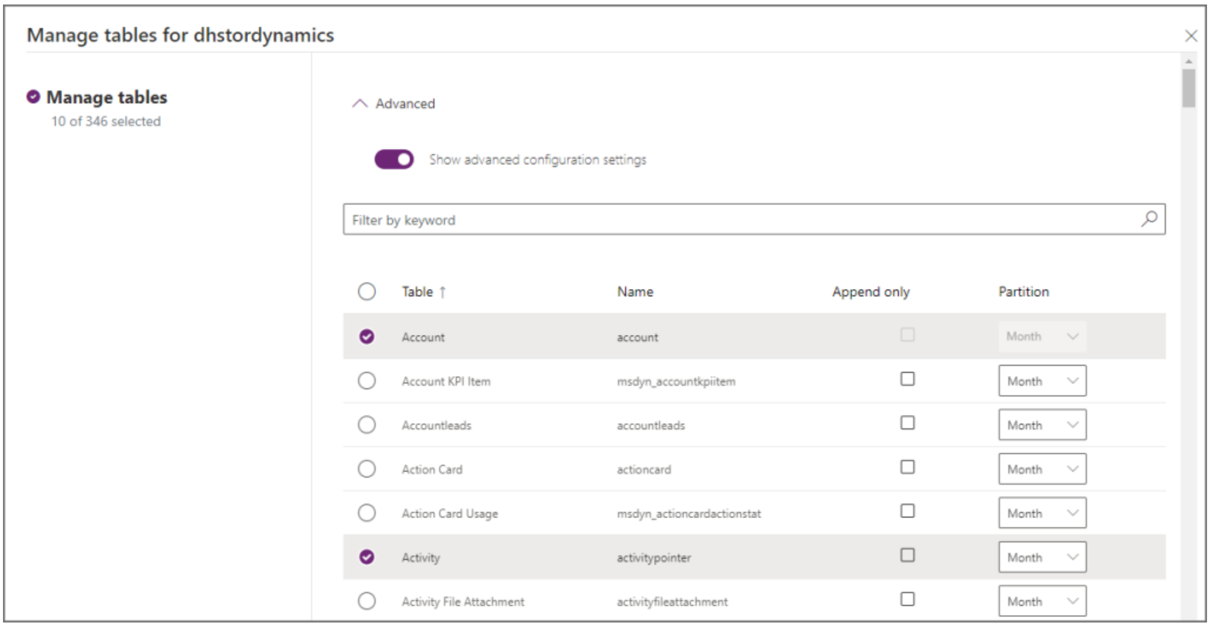
Azure Synapse Link for Dataverse: Understanding Advanced Configuration Settings

In this section we'll look at the **advanced configuration settings** available when selecting the Dataverse tables we would like to export via Synapse Link.

Advanced Configuration Settings

When configuring Synapse Link for Dataverse in the Power Apps portal, there is an option **Show advanced configuration settings** which when enabled, shows 2 extra column options:

- Append Only
- Partition



Manage tables for dhstordynamics

Manage tables
10 of 346 selected

Advanced

Show advanced configuration settings

Filter by keyword

○	Table ↑	Name	Append only	Partition
●	Account	account	□	Month
○	Account KPI Item	msdyn_accountkpiitem	□	Month
○	Accountleads	accountleads	□	Month
○	Action Card	actioncard	□	Month
○	Action Card Usage	msdyn_actioncardactionstat	□	Month
●	Activity	activitypointer	□	Month
○	Activity File Attachment	activityfileattachment	□	Month

Advanced configuration settings

Partition

We'll look at Partition first as it's also relevant to the Append Only section. The partition option specifies how the CSV files will be partitioned in the Azure Data Lake. There are 2 options available:

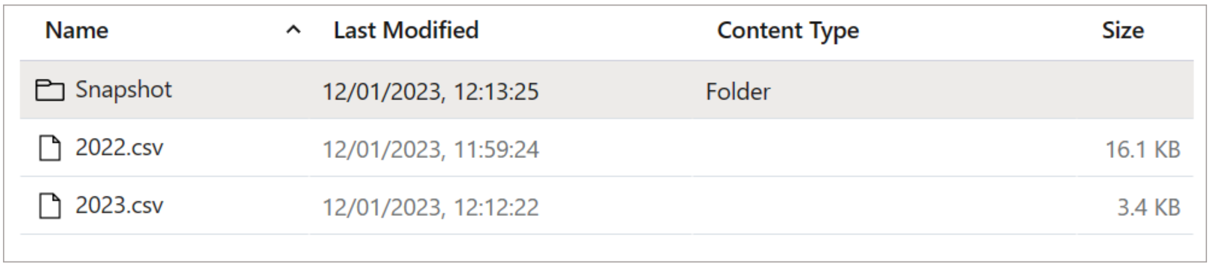
- **Month** (example 2021-12.csv, 2022-12.csv)
- **Year** (example 2021.csv, 2022.csv)

The partitioning is based on the **createdOn** date column in the source Dataverse table, the date the record was created in the dataverse table dictates which partition file it is written to.

Note that when **Append Only** is enabled, the Partition value will default to **Year** and cannot be changed.

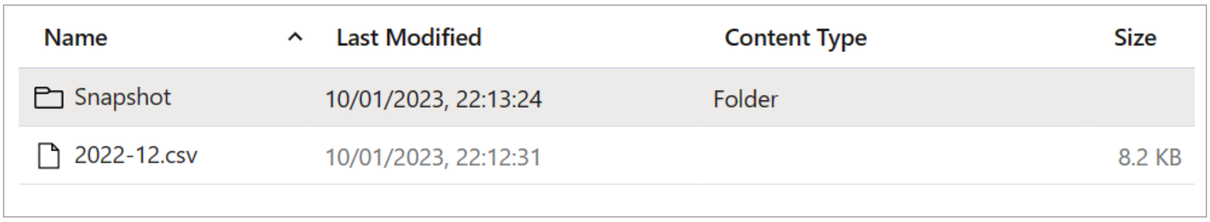
And yes, the only option now is CSV files… Parquet support is coming Q1 2023 apparently, I am on the case with this as it'll be a great option to have.

Partition by Year



Name	Last Modified	Content Type	Size
Snapshot	12/01/2023, 12:13:25	Folder	
2022.csv	12/01/2023, 11:59:24		16.1 KB
2023.csv	12/01/2023, 12:12:22		3.4 KB

Partition by Month



Name	Last Modified	Content Type	Size
Snapshot	10/01/2023, 22:13:24	Folder	
2022-12.csv	10/01/2023, 22:12:31		8.2 KB

Append Only vs In Place Update

To understand **Append Only** we'll also look at **In Place Update** which is the default setting. At its most basic, Append Only does not delete any data from Azure Data Lake when it is deleted from the dataverse environment. Rather, it is marked as deleted. When data is updated, the original row is kept in the Data Lake and a new row with the updated data is added. We'll look at this with examples below.

Append Only is the recommended choice if you wish to use point-in-time data for historical analysis or as part of a Data Warehousing loading process (think Slowly Changing Dimensions). As an aside, the snapshot data does contain changed data, but this is not readily accessible.

Note that when **Append Only** is enabled, the Partition value will default to **Year** and cannot be changed.

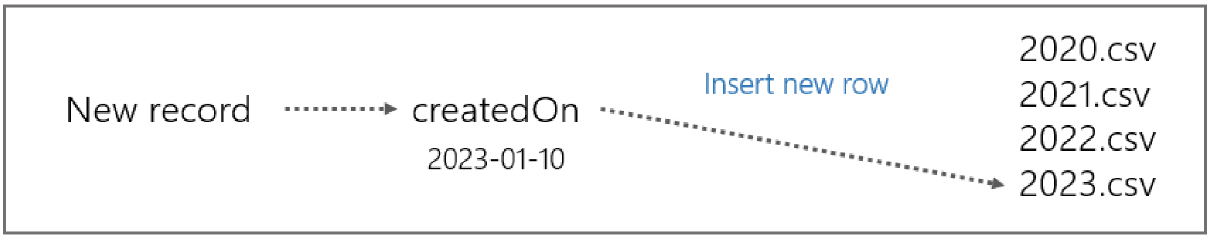
Also note that **Append only** is also the default setting for Dataverse tables that don't have a **createdOn** value.

In Place Update (Default)

Let's now work through examples for new, updated, and deleted data. For In Place Updates we'll show the logic in the diagrams, for Append Only we'll dive deeper into the data itself.

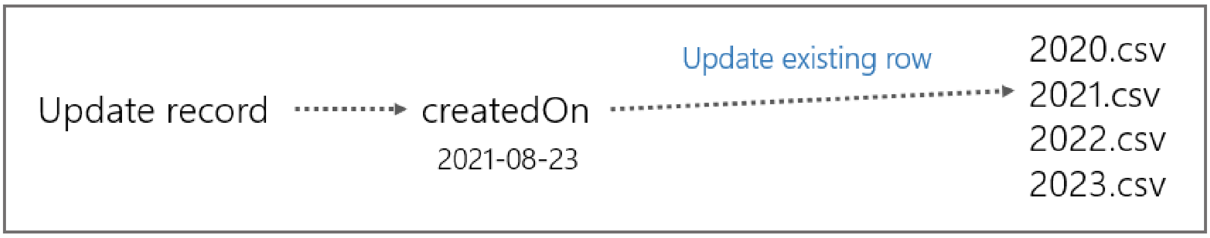
Insert New Row

When a new record is added to the Dataverse, the **createdOn** date will determine which yearly partitioned file it is written to. The record is then **inserted** into the appropriate file. The examples below are based on the partition being set to Year.



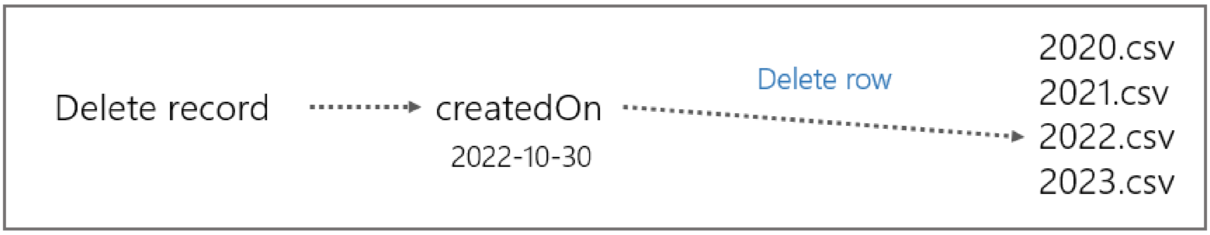
Update Existing Record

When an existing record is updated, the createdOn date determines which file will be scanned for an existing record. It is then **updated** if the record exists.



Delete Existing Record

When a record is deleted, the createdOn date determines which file will be scanned for an existing record. It is then **permanently deleted** from the file.

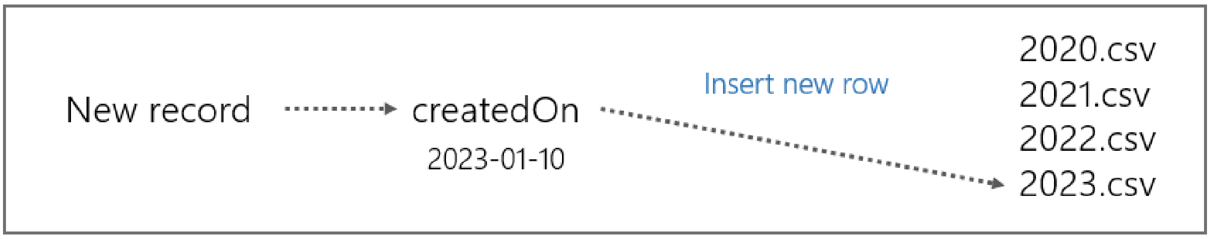


Append Only

Let’s now look at the Append Only feature, with data examples for inserting, updating, and deleting data in the dataverse tables. We’ll be using the **contact** table as an example.

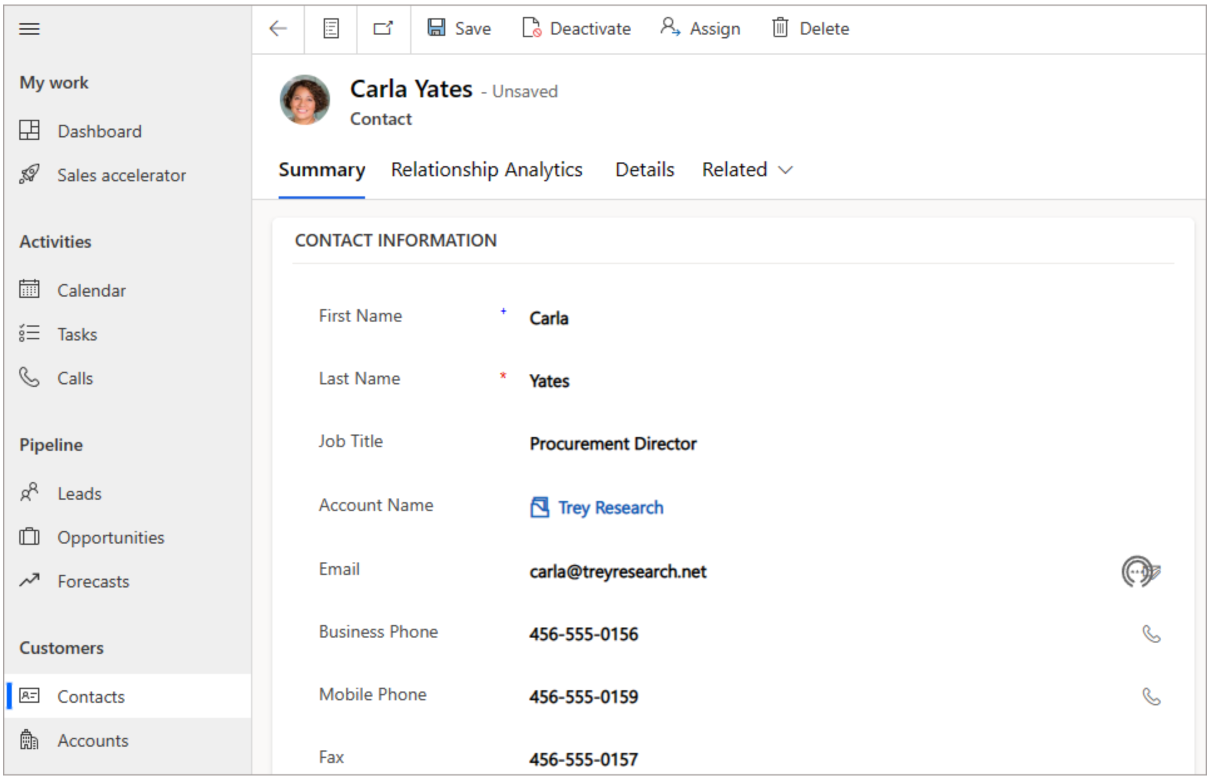
Insert New Row

This is the same as In Place Update, when a new record is added to the Dataverse, the **createdOn** date will determine which yearly partitioned file it is written to. The record is then **inserted** into the appropriate file.



Update Existing Record

If we update an existing **contact** record in Dynamics 365 Sales, E.G. update Carla Yates Job Title from **Procurement Manager** to **Procurement Director**, the new row plus the existing row is written to the end of the relevant partition file based on original createdOn date. The **modifiedon** datetime will indicate when the change occurred in the source system so this can be used in any downstream ETL/ELT solution. The original row will not have its values updated; the new row will contain the relevant data including the modifiedon datetime.



Carla Yates - Unsaved
Contact

CONTACT INFORMATION

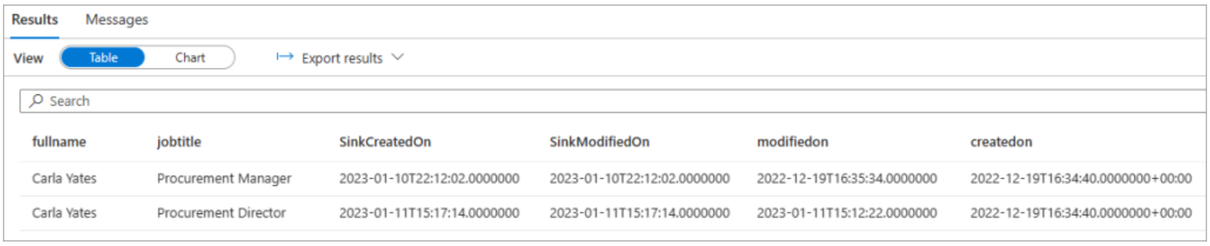
First Name *	Carla
Last Name *	Yates
Job Title	Procurement Director
Account Name	Trey Research
Email	carla@treyresearch.net
Business Phone	456-555-0156
Mobile Phone	456-555-0159
Fax	456-555-0157

Updating a record in Dynamics 365 Sales

Let's now query the **contacts** table in the Lake Database using Synapse Serverless SQL Pools.

```sql
SELECT
		fullname,
		jobtitle,
		SinkCreatedOn,
		SinkModifiedOn,
		modifiedon,
		createdon
FROM contact
WHERE fullname = 'Carla Yates';
```

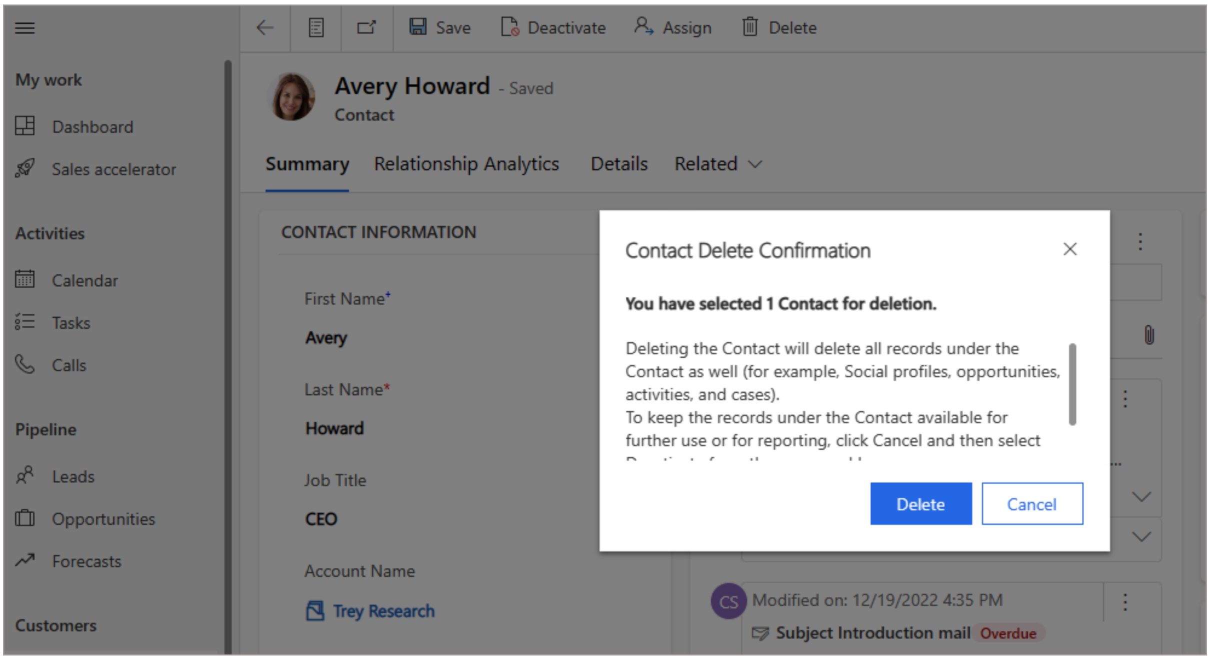
We can see the results below, we have 2 rows: the existing unchanged row, and the new “updated” row. We can use the modifedon column in any data loading/etl/elt process to determine changed data. Also, we can use the **Id** column (not in the example below) to correlate the rows together. We’ll see the use of Id later.



fullname	jobtitle	SinkCreatedOn	SinkModifiedOn	modifiedon	createdon
Carla Yates	Procurement Manager	2023-01-10T22:12:02.0000000	2023-01-10T22:12:02.0000000	2022-12-19T16:35:34.0000000	2022-12-19T16:34:40.0000000+00:00
Carla Yates	Procurement Director	2023-01-11T15:17:14.0000000	2023-01-11T15:17:14.0000000	2023-01-11T15:12:22.0000000	2022-12-19T16:34:40.0000000+00:00

Delete Existing Record

If we hard-delete (not just disable) the record from Dataverse/Dynamics, then we still see the original row unchanged in the partitioned file and we also see a **new row added** with **IsDelete** set to **true**. The rows are correlated on the **Id** column. Note that most of the column values for this new record are now NULL. This can now be used in any downstream ETL/ELT solution. BTW the official documentation states the column is called isDeleted, but during my testing the column is called IsDelete.



Avery Howard - Saved

Contact

CONTACT INFORMATION

First Name*

Avery

Last Name*

Howard

Job Title

CEO

Account Name

Trey Research

Contact Delete Confirmation

You have selected 1 Contact for deletion.

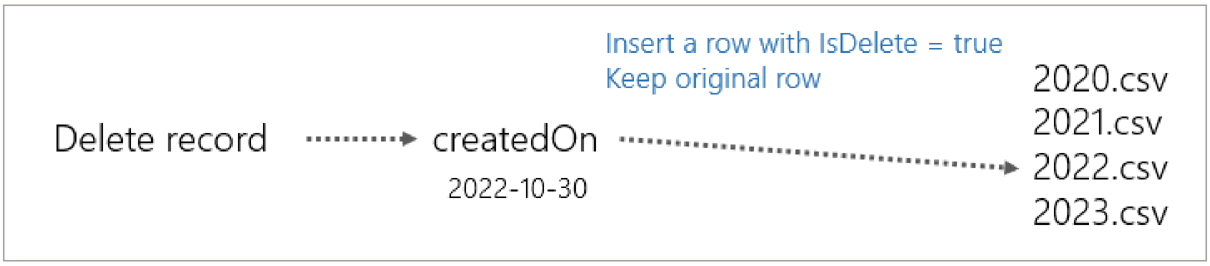
Deleting the Contact will delete all records under the Contact as well (for example, Social profiles, opportunities, activities, and cases).
To keep the records under the Contact available for further use or for reporting, click Cancel and then select

Modified on: 12/19/2022 4:35 PM

Subject Introduction mail Overdue

Deleting a record from Dynamics 365 Sales

The logic for an Append Only delete is to keep the original row and insert a new row with the IsDelete flag set to **true**.



Delete record

createdOn

2022-10-30

Insert a row with IsDelete = true

Keep original row

2020.csv

2021.csv

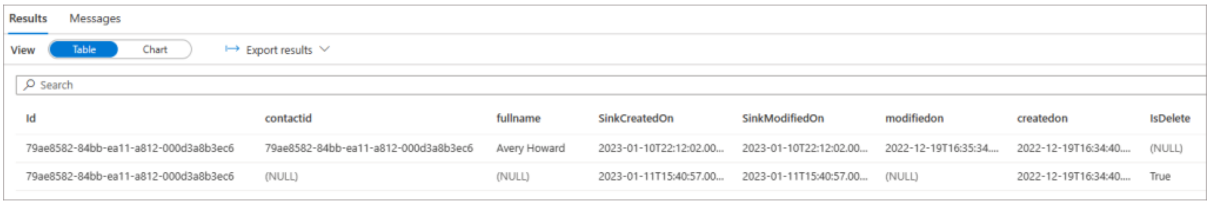
2022.csv

2023.csv

If we run a SELECT using Serverless SQL Pools for the specific **Id**, we can see the results below.

```
SELECT
        Id,
        contactid,
        fullname,
        SinkCreatedOn,
        SinkModifiedOn,
        modifiedon,
        createdon,
        IsDelete
FROM contact
WHERE id = '79ae8582-84bb-ea11-a812-000d3a8b3ec6';
```

We can see the original unchanged row plus the new row with NULL for most attributes and with IsDelete set to true.

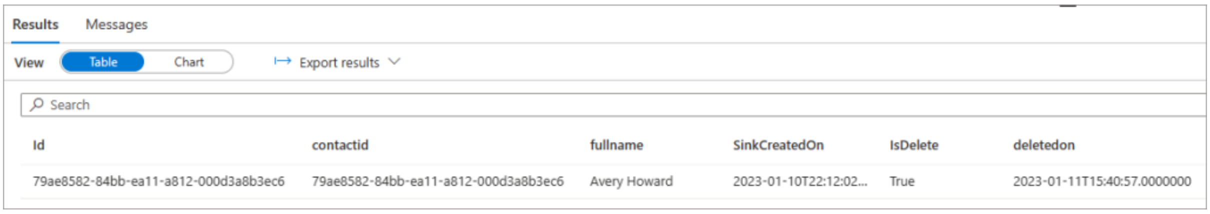


Id	contactid	fullname	SinkCreatedOn	SinkModifiedOn	modifiedon	createdon	IsDelete
79ae8582-84bb-ea11-a812-000d3a8b3ec6	79ae8582-84bb-ea11-a812-000d3a8b3ec6	Avery Howard	2023-01-10T22:12:02.00...	2023-01-10T22:12:02.00...	2022-12-19T16:35:34....	2022-12-19T16:34:40....	(NULL)
79ae8582-84bb-ea11-a812-000d3a8b3ec6	(NULL)	(NULL)	2023-01-11T15:40:57.00...	2023-01-11T15:40:57.00...	(NULL)	2022-12-19T16:34:40....	True

Let’s write a SQL query to get the deleted row and join back to the original row and flag as deleted.

```
;WITH deletedrows
AS
(
        SELECT
                Id,
                SinkCreatedOn,
                IsDelete
        FROM contact
        WHERE IsDelete = 'True'
)
SELECT
        c.Id,
        c.contactid,
        c.fullname,
        c.SinkCreatedOn,
        d.IsDelete,
        d.SinkCreatedOn AS deletedon
FROM contact c
LEFT JOIN deletedrows d ON c.Id = d.Id
WHERE c.IsDelete IS NULL AND c.fullname = 'Avery Howard'
ORDER BY c.Id;
```

We’ve now got a single row in the results with the original values plus the IsDelete column and the derived **deleteon** column we created using the SinkCreatedOn value from the new “deleted” row. This could be useful for any archiving/data loading process.



Id	contactid	fullname	SinkCreatedOn	IsDelete	deletedon
79ae8582-84bb-ea11-a812-000d3a8b3ec6	79ae8582-84bb-ea11-a812-000d3a8b3ec6	Avery Howard	2023-01-10T22:12:02...	True	2023-01-11T15:40:57.0000000

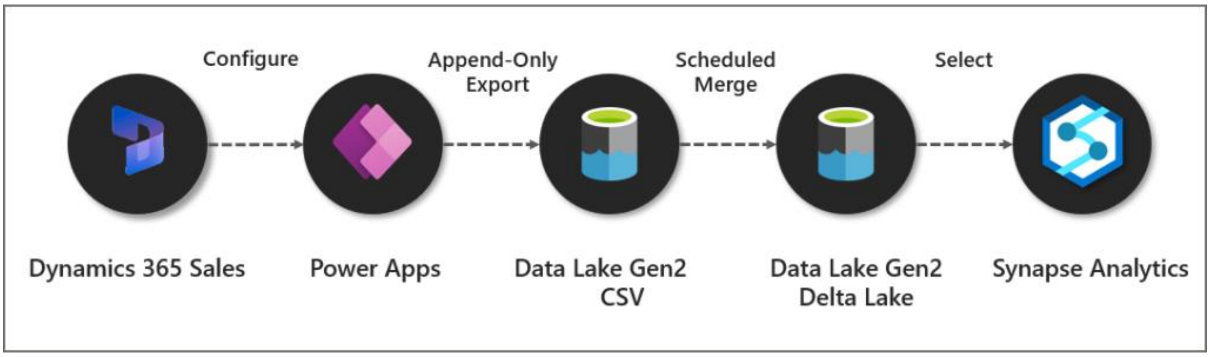
Synapse Link for Dataverse: Exporting to Delta Lake

Microsoft recently released (~April 2023) the ability to configure **Synapse Link for Dataverse** and use **Delta Lake** as the export format. **Delta Lake** is a data and transaction storage file format very popular in Lakehouse implementations and is becoming a very popular format that enables decoupling storage with compute. More info here

The official documentation is here.

Process Overview

There are several moving parts involved in the process when setting up the sync. This involves the initial configuration in the Power Apps portal to configure the Synapse Link, then data is exported from the Dataverse into the Data Lake in CSV format, then it's merged into a Delta Lake folder (a folder for each table). The Delta Lake folder is then queryable using Synapse Serverless SQL Pools and Spark pools as tables are created in a Lake Database.



Cost

Let's raise this first, this feature is chargeable as you need to create a Spark Pool (cluster) within a Synapse Analytics workspace. Now, I'm always keen to identify **cost** vs **value**. Something that costs money doesn't necessarily mean it doesn't provide value, and vice-verse.

I saw a daily cost of £1.23 even though no data had changed in the source Dynamics 365 Sales instance. This cost was due to the daily Delta maintenance that is performed (file compaction and removing of old files) by running a Spark job.

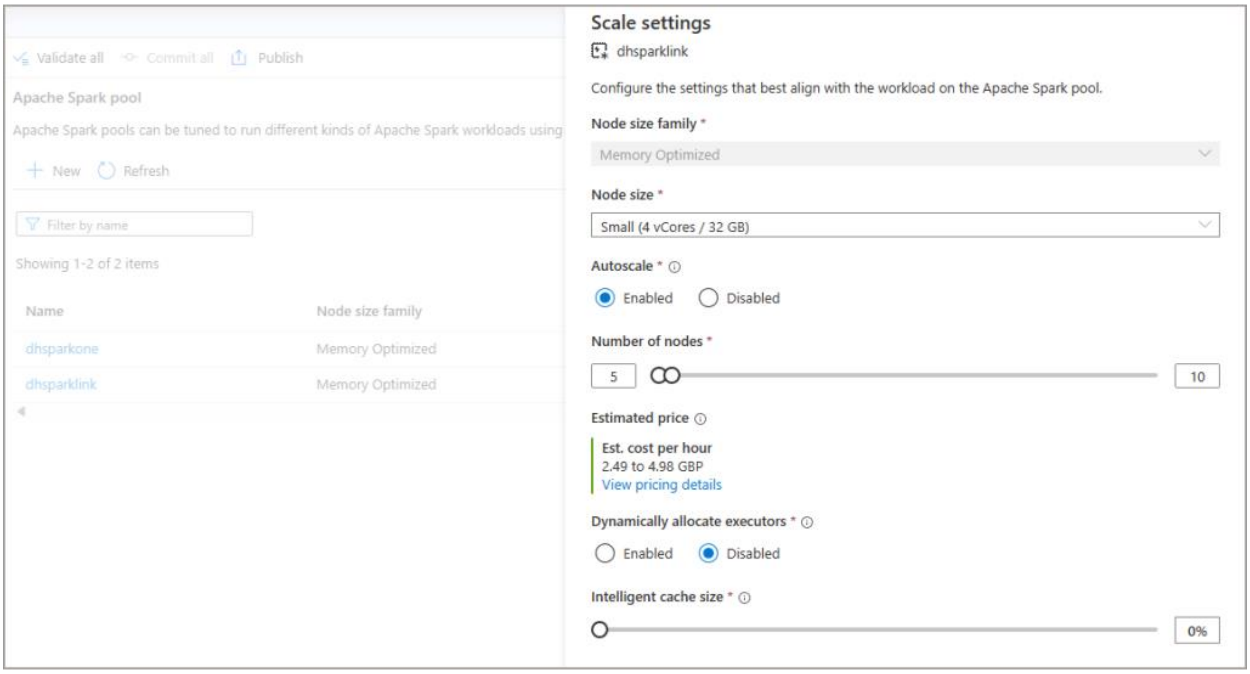
Walkthrough

In this section it's assumed that Dynamics Sales is already configured, a Synapse Analytics workspace is provisioned, and the user has appropriate permissions.

Setup Spark Pool

The first thing to do is set-up a Spark pool in an Azure Synapse Analytics workspace that will be used in the syncing process.

- Login to the Synapse Analytics workspace
- Click **Manage** > **Apache Spark Pools**
- Click **New** and enter the relevant information
- Make sure in **Additional Settings** that the Spark version is set to 3.1
- Below is an image of the configuration I used successfully.



Apache Spark pool

Apache Spark pools can be tuned to run different kinds of Apache Spark workloads using

Showing 1-2 of 2 items

Name	Node size family
dhsparkone	Memory Optimized
dhsparklink	Memory Optimized

Scale settings

dhsparklink

Configure the settings that best align with the workload on the Apache Spark pool.

Node size family *
Memory Optimized

Node size *
Small (4 vCores / 32 GB)

Autoscale *
● Enabled ○ Disabled

Number of nodes *
5 – 10

Estimated price
Est. cost per hour
2.49 to 4.98 GBP
View pricing details

Dynamically allocate executors *
○ Enabled ● Disabled

Intelligent cache size *
0%

I did encounter an error when I set the number of nodes too low, the Synapse Link process really does want a minimum node number of 5 here.

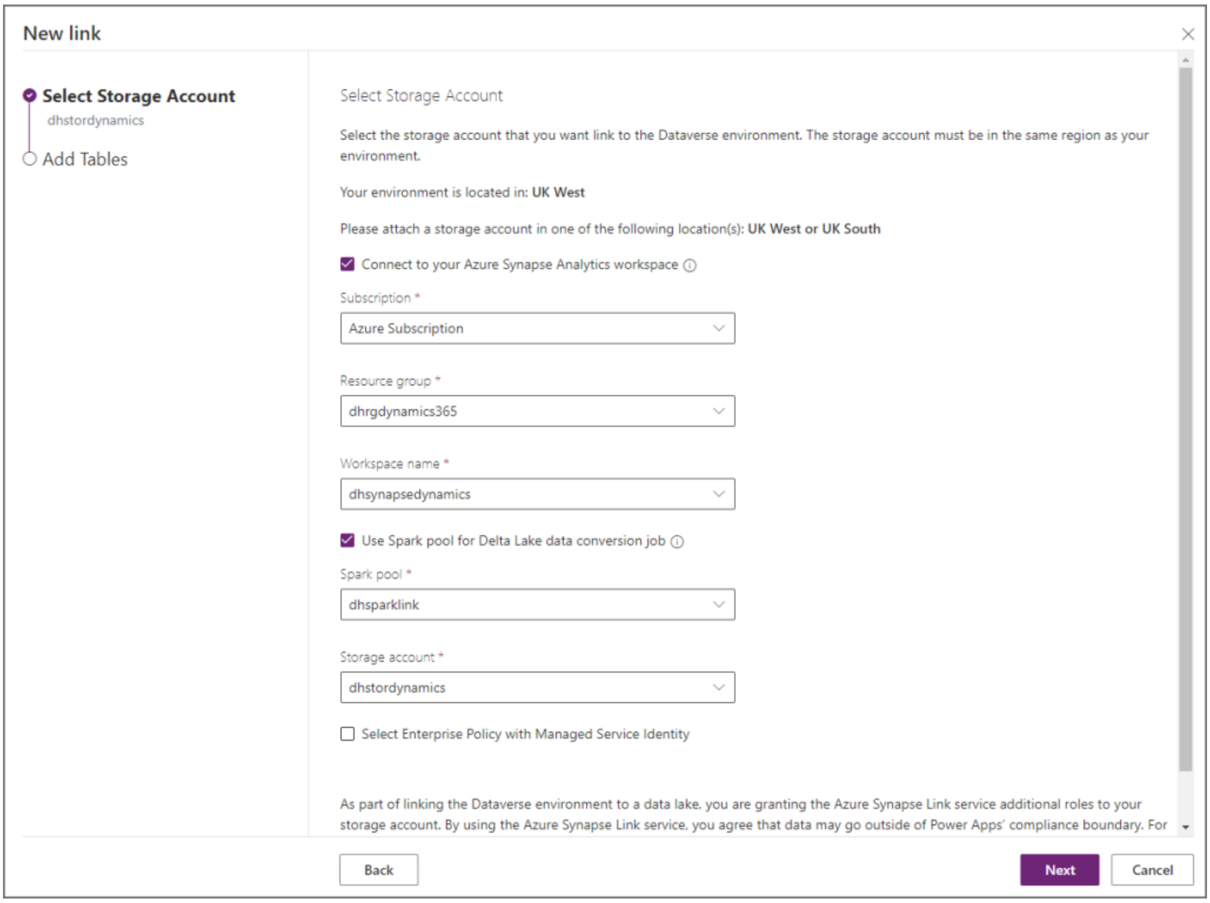
Synapse Link Configuration

Now that the Spark Pool has been setup, let's head on over to Power Apps and configure the tables we want to export.

Please note the location of the Synapse Link item may not be immediately available on the left side menu. Hover over **Discover** then click the **Discover all** button. You should then see **Synapse Link** under **Data Management**, you can then pin the item to the menu.

- Login to [Power Apps]()
- Select the relevant **environment** from the top-right menu
- On the left menu, select **Dataverse** -> **Azure Synapse Link**
- You will need to append **?athena.deltaLake=true** to the end of the current URL.
  E.G. **https://make.powerapps.com/environments/<environment_guid>/exporttodatalake?athena.deltaLake=true**
- Click **New Link** and enter the following information:
  - Enable **Connect to your Azure Synapse Analytics workspace**
  - **Subscription**: Select the appropriate subscription
  - **Resource group**: Select the resource group the Synapse workspace is in
  - **Workspace Name**: Select the specific Synapse workspace
  - Enable **Use Spark pool for Delta Lake data conversion job**
  - Then in the **Spark Pool** drop-down, select the Spark pool created earlier
  - **Storage account**: Select the appropriate storage account to use
- Click **Next** and select the tables you want to syncronise.

Note that we cannot configure any of the properties like Append Only, Partition etc. We can configure the **Time Interval** which will export the data to the data lake within specific folder datetimes (btw this is not a setting to dictate when data is exported to the Data Lake as that is "near real-time", just the folder structure).



New link

- Select Storage Account
  dhstordynamics
- Add Tables

Select Storage Account

Select the storage account that you want link to the Dataverse environment. The storage account must be in the same region as your environment.

Your environment is located in: **UK West**

Please attach a storage account in one of the following location(s): **UK West or UK South**

☑ Connect to your Azure Synapse Analytics workspace

Subscription *: Azure Subscription

Resource group *: dhrgdynamics365

Workspace name *: dhsynapsedynamics

☑ Use Spark pool for Delta Lake data conversion job

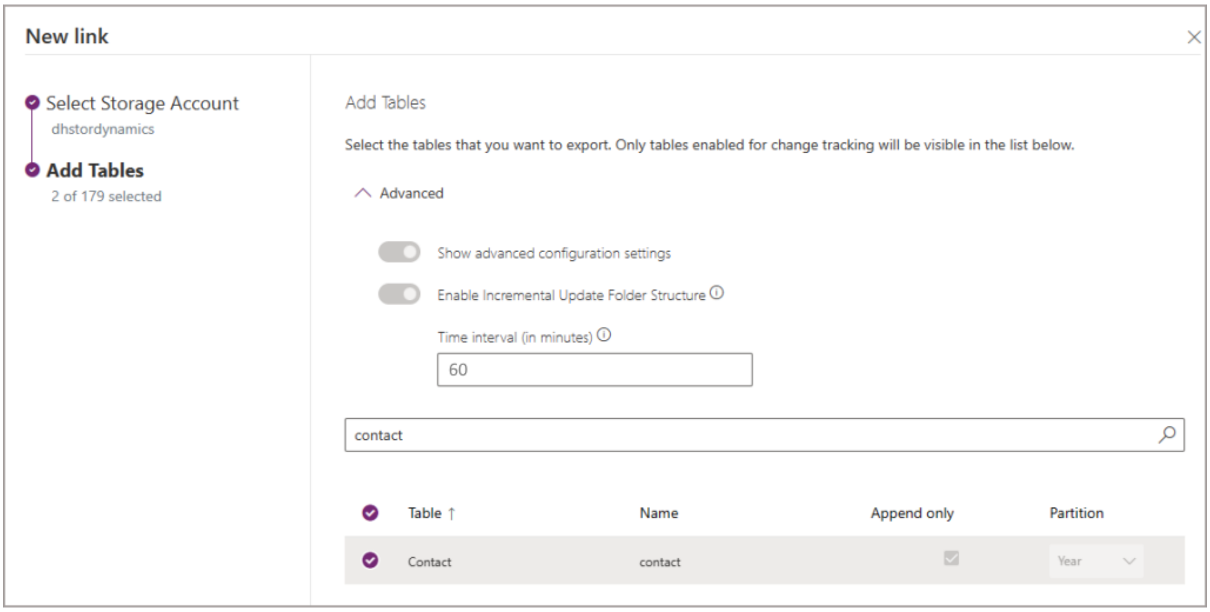
Spark pool *: dhsparklink

Storage account *: dhstordynamics

□ Select Enterprise Policy with Managed Service Identity

As part of linking the Dataverse environment to a data lake, you are granting the Azure Synapse Link service additional roles to your storage account. By using the Azure Synapse Link service, you agree that data may go outside of Power Apps' compliance boundary. For

In the configuration below, I've set the incremental folder structure time interval to be 60 minutes. This groups up all the changes into folders which are timestamped appropriately (e.g. the folders will be generated every 60 minutes). As I explained earlier, this has nothing to do with when the data is exported, just the naming of the folders in the data lake.



New link

- Select Storage Account
  dhstordynamics
- Add Tables
  2 of 179 selected

Add Tables

Select the tables that you want to export. Only tables enabled for change tracking will be visible in the list below.

Advanced

Show advanced configuration settings: On

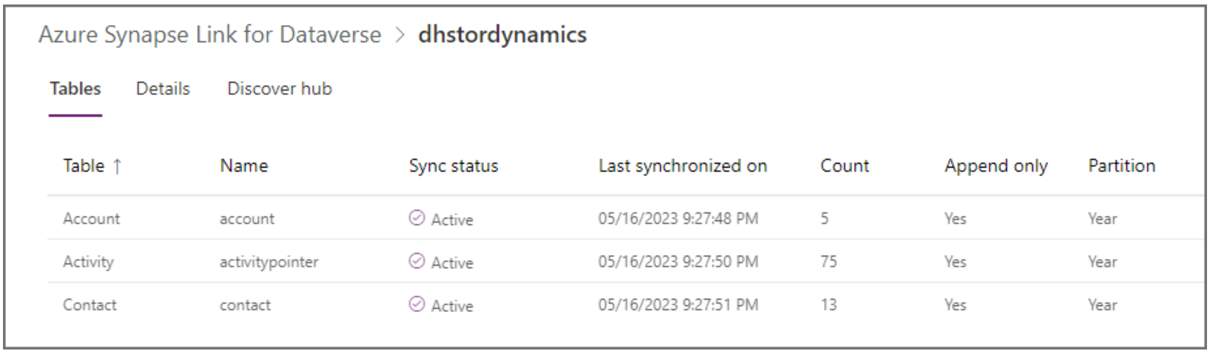
Enable Incremental Update Folder Structure: On

Time interval (in minutes): 60

contact

	Table ↑	Name	Append only	Partition
✓	Contact	contact	☑	Year

After the tables have been setup, you'll see a table showing the sync status.



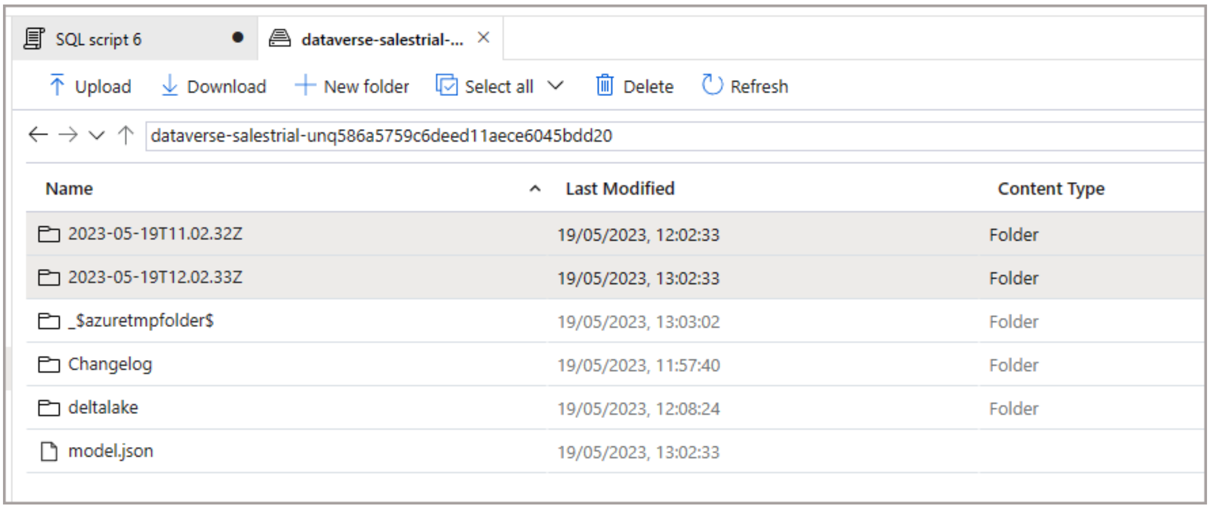
Azure Synapse Link for Dataverse > dhstordynamics

Table ↑	Name	Sync status	Last synchronized on	Count	Append only	Partition
Account	account	✓ Active	05/16/2023 9:27:48 PM	5	Yes	Year
Activity	activitypointer	✓ Active	05/16/2023 9:27:50 PM	75	Yes	Year
Contact	contact	✓ Active	05/16/2023 9:27:51 PM	13	Yes	Year

CSV Export

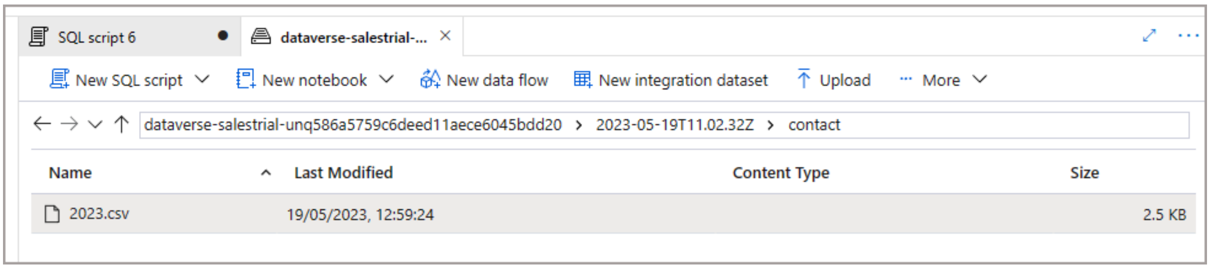
Now if we switch over to the Data Lake storage account that was used when setting up the sync, we see a new container with the Dynamics environment GUID and a set of folders. When data is being exported from the Dataverse into the Data Lake, it will be stored in the **datetime** folders in CSV format. It's worth noting now that once the Delta Lake merging process has been completed, the CSV files are removed automatically.

NB: the exported folder structure is timestamped in 60 minute intervals.



dataverse-salestrial-unq586a5759c6deed11aece6045bdd20

Name	Last Modified	Content Type
2023-05-19T11.02.32Z	19/05/2023, 12:02:33	Folder
2023-05-19T12.02.33Z	19/05/2023, 13:02:33	Folder
_$azuretmpfolder$	19/05/2023, 13:03:02	Folder
Changelog	19/05/2023, 11:57:40	Folder
deltalake	19/05/2023, 12:08:24	Folder
model.json	19/05/2023, 13:02:33	

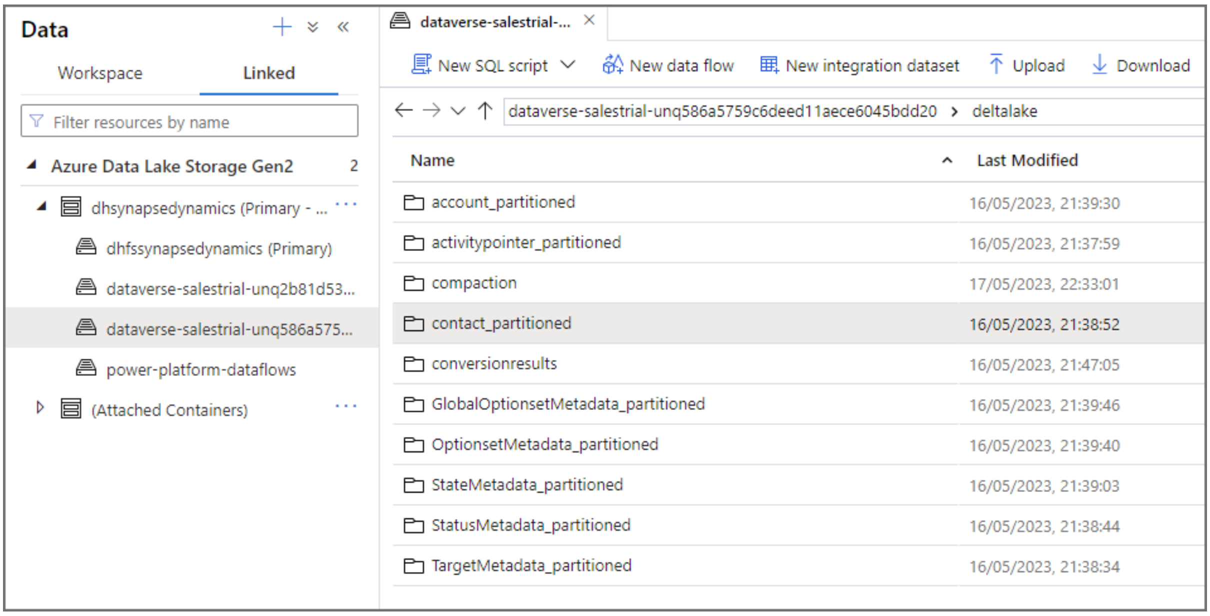


dataverse-salestrial-unq586a5759c6deed11aece6045bdd20 > 2023-05-19T11.02.32Z > contact

Name	Last Modified	Content Type	Size
2023.csv	19/05/2023, 12:59:24		2.5 KB

Delta Lake

It takes a few minutes but then in the root container, there will be a folder called **deltalake**. This contains all the tables that were setup to be synchronised (also includes metadata tables too). Each folder is the table itself and inside each folder will be the parquet files and also the Delta log.

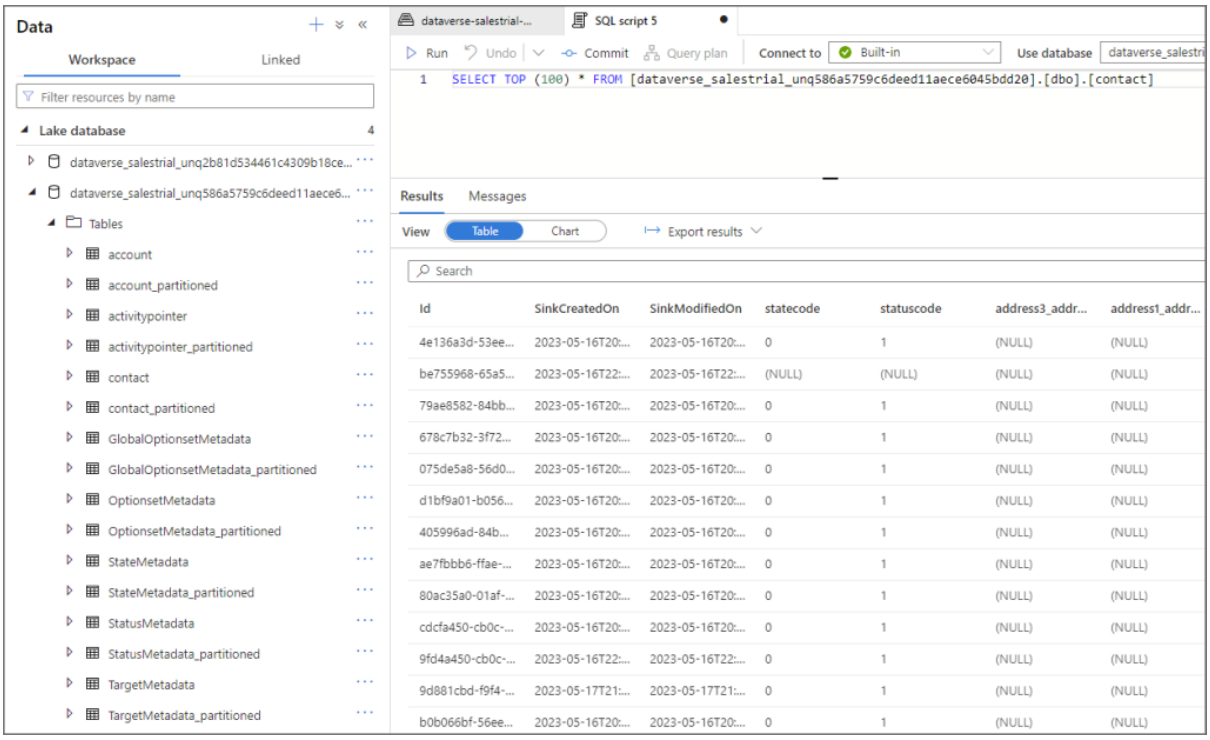


dataverse-salestrial-unq586a5759c6deed11aece6045bdd20 › deltalake

Name	Last Modified
account_partitioned	16/05/2023, 21:39:30
activitypointer_partitioned	16/05/2023, 21:37:59
compaction	17/05/2023, 22:33:01
contact_partitioned	16/05/2023, 21:38:52
conversionresults	16/05/2023, 21:47:05
GlobalOptionsetMetadata_partitioned	16/05/2023, 21:39:46
OptionsetMetadata_partitioned	16/05/2023, 21:39:40
StateMetadata_partitioned	16/05/2023, 21:39:03
StatusMetadata_partitioned	16/05/2023, 21:38:44
TargetMetadata_partitioned	16/05/2023, 21:38:34

Querying the Delta Lake Tables

Once the Delta Lake folders have been created, we should see a new database in the **Lake Database** area in Synapse. We can now query the tables using either Serverless SQL Pools or Spark pools. It’s worth noting that querying via Serverless SQL Pools does not allow you to query “point-in-time” as per Delta Lake functionality, you’ll just get the latest version of the data. We can use time-travel in Spark pools though.

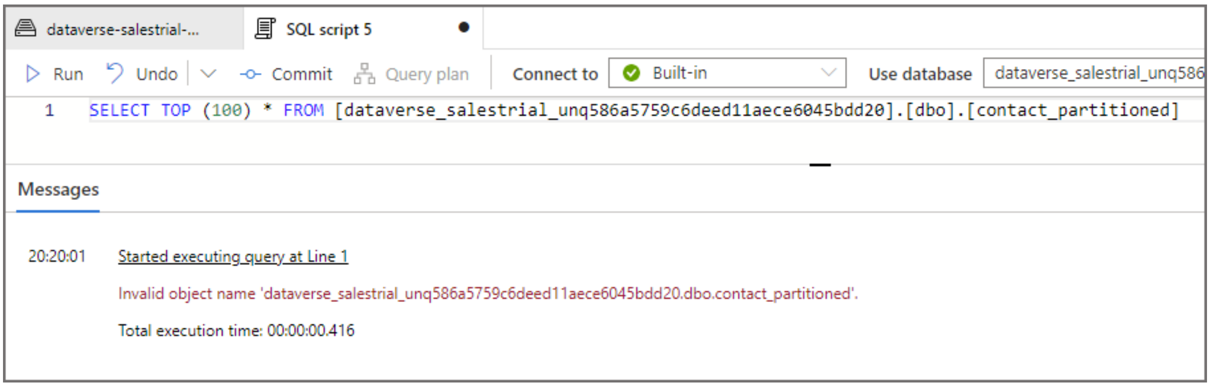


```sql
SELECT TOP (100) * FROM [dataverse_salestrial_unq586a5759c6deed11aece6045bdd20].[dbo].[contact]
```

Id	SinkCreatedOn	SinkModifiedOn	statecode	statuscode	address3_addr...	address1_addr...
4e136a3d-53ee...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)
be755968-65a5...	2023-05-16T22:...	2023-05-16T22:...	(NULL)	(NULL)	(NULL)	(NULL)
79ae8582-84bb...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)
678c7b32-3f72...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)
075de5a8-56d0...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)
d1bf9a01-b056...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)
405996ad-84b...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)
ae7fbbb6-ffae-...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)
80ac35a0-01af-...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)
cdcfa450-cb0c-...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)
9fd4a450-cb0c-...	2023-05-16T22:...	2023-05-16T22:...	0	1	(NULL)	(NULL)
9d881cbd-f9f4-...	2023-05-17T21:...	2023-05-17T21:...	0	1	(NULL)	(NULL)
b0b066bf-56ee...	2023-05-16T20:...	2023-05-16T20:...	0	1	(NULL)	(NULL)

I've noticed something confusing with the table names in the Lake database. As per standard process, there will be 2 tables created for each source table. E.G. for the contact table, there will be **contact** and **contact_partitioned** (reason in having 2 tables is the base table is near real-time and the partitioned table has a longer update interval to avoid any potential file locking).

But here, only the **contact** table can be queried as **contact_partitioned** generates an error stating it's an invalid object name.

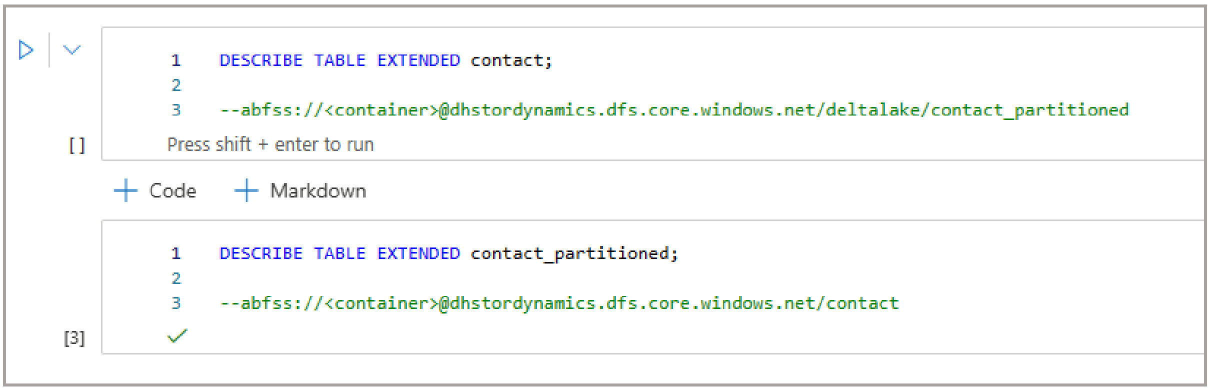


```sql
SELECT TOP (100) * FROM [dataverse_salestrial_unq586a5759c6deed11aece6045bdd20].[dbo].[contact_partitioned]
```

```
20:20:01    Started executing query at Line 1
            Invalid object name 'dataverse_salestrial_unq586a5759c6deed11aece6045bdd20.dbo.contact_partitioned'.
            Total execution time: 00:00:00.416
```

When I look at the table metadata using Spark I can see that the **contact** table location actually looks at the **deltalake\contact_partitioned** folder…that's a little confusing. Also, the data lake location for **contact_partitioned** which is **/contact** doesn't actually exist.

I've reached out to Microsoft for clarification on this.



```
DESCRIBE TABLE EXTENDED contact;

--abfss://<container>@dhstordynamics.dfs.core.windows.net/deltalake/contact_partitioned
```

[ ] Press shift + enter to run

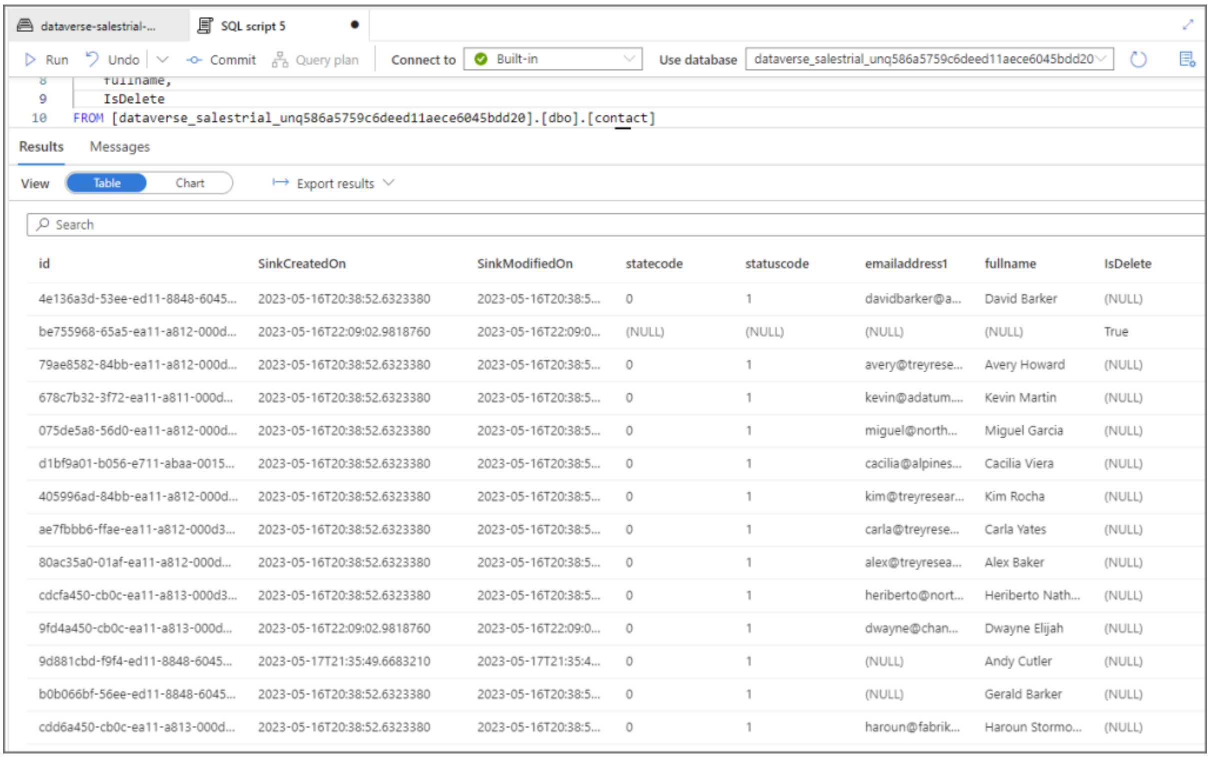
```
DESCRIBE TABLE EXTENDED contact_partitioned;

--abfss://<container>@dhstordynamics.dfs.core.windows.net/contact
```

[3] ✓

Append Only

As the sync is automatically configured to use Append-Only, what does that mean? Well, it means the export from Dynamics does not hard-delete any records. If data is deleted or updated in the source, those deletions and updates do not overwrite the destination data in the data lake. E.G. if a record is deleted in Dynamics, then you'll see **True** in the **IsDelete** column.



```
    fullname,
    IsDelete
FROM [dataverse_salestrial_unq586a5759c6deed11aece6045bdd20].[dbo].[contact]
```

id	SinkCreatedOn	SinkModifiedOn	statecode	statuscode	emailaddress1	fullname	IsDelete
4e136a3d-53ee-ed11-8848-6045...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	davidbarker@a...	David Barker	(NULL)
be755968-65a5-ea11-a812-000d...	2023-05-16T22:09:02.9818760	2023-05-16T22:09:0...	(NULL)	(NULL)	(NULL)	(NULL)	True
79ae8582-84bb-ea11-a812-000d...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	avery@treyrese...	Avery Howard	(NULL)
678c7b32-3f72-ea11-a811-000d...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	kevin@adatum....	Kevin Martin	(NULL)
075de5a8-56d0-ea11-a812-000d...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	miguel@north...	Miguel Garcia	(NULL)
d1bf9a01-b056-e711-abaa-0015...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	cacilia@alpines...	Cacilia Viera	(NULL)
405996ad-84bb-ea11-a812-000d...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	kim@treyresear...	Kim Rocha	(NULL)
ae7fbbb6-ffae-ea11-a812-000d3...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	carla@treyrese...	Carla Yates	(NULL)
80ac35a0-01af-ea11-a812-000d...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	alex@treyresea...	Alex Baker	(NULL)
cdcfa450-cb0c-ea11-a813-000d3...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	heriberto@nort...	Heriberto Nath...	(NULL)
9fd4a450-cb0c-ea11-a813-000d...	2023-05-16T22:09:02.9818760	2023-05-16T22:09:0...	0	1	dwayne@chan...	Dwayne Elijah	(NULL)
9d881cbd-f9f4-ed11-8848-6045...	2023-05-17T21:35:49.6683210	2023-05-17T21:35:4...	0	1	(NULL)	Andy Cutler	(NULL)
b0b066bf-56ee-ed11-8848-6045...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	(NULL)	Gerald Barker	(NULL)
cdd6a450-cb0c-ea11-a813-000d...	2023-05-16T20:38:52.6323380	2023-05-16T20:38:5...	0	1	haroun@fabrik...	Haroun Stormo...	(NULL)

Time Travel

As I said earlier, Serverless SQL Pools doesn't have an official way of time-travel in Delta Lake so when querying a Delta folder you'll always get the latest data. But we are able to use Spark to time-travel in the Dynamics data that's synced.

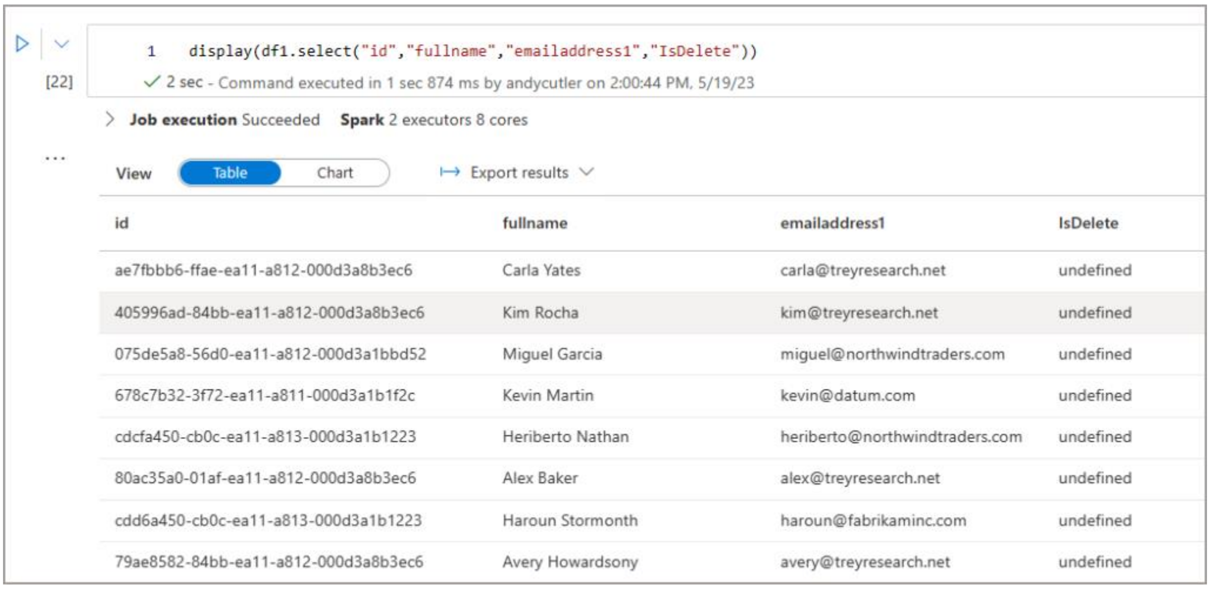
```
df1 = spark.read \

.format("delta") \

.option("timestampAsOf", "2023-05-19 11:09:30.942") \

.load("abfss://<container>@dhstordynamics.dfs.core.windows.net/deltalake/contact_partitioned") \

.show()
```



```
display(df1.select("id","fullname","emailaddress1","IsDelete"))
```

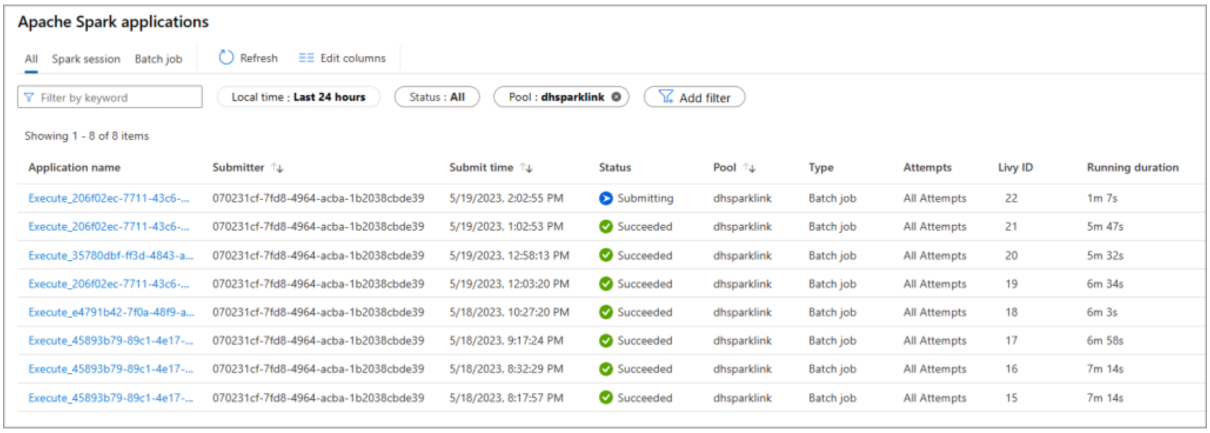
✓ 2 sec - Command executed in 1 sec 874 ms by andycutler on 2:00:44 PM, 5/19/23

Job execution Succeeded Spark 2 executors 8 cores

id	fullname	emailaddress1	IsDelete
ae7fbbb6-ffae-ea11-a812-000d3a8b3ec6	Carla Yates	carla@treyresearch.net	undefined
405996ad-84bb-ea11-a812-000d3a8b3ec6	Kim Rocha	kim@treyresearch.net	undefined
075de5a8-56d0-ea11-a812-000d3a1bbd52	Miguel Garcia	miguel@northwindtraders.com	undefined
678c7b32-3f72-ea11-a811-000d3a1b1f2c	Kevin Martin	kevin@datum.com	undefined
cdcfa450-cb0c-ea11-a813-000d3a1b1223	Heriberto Nathan	heriberto@northwindtraders.com	undefined
80ac35a0-01af-ea11-a812-000d3a8b3ec6	Alex Baker	alex@treyresearch.net	undefined
cdd6a450-cb0c-ea11-a813-000d3a1b1223	Haroun Stormonth	haroun@fabrikaminc.com	undefined
79ae8582-84bb-ea11-a812-000d3a8b3ec6	Avery Howardsony	avery@treyresearch.net	undefined

Monitoring

We can monitor the execution of the Spark jobs which perform the merging and also the daily maintenance in the Synapse workspace by clicking **Monitor** on the main Synapse menu and clicking **Apache Spark Applications**. This shows when the Spark jobs ran and the duration of the jobs.



Showing 1 - 8 of 8 items

Application name	Submitter	Submit time	Status	Pool	Type	Attempts	Livy ID	Running duration
Execute_206f02ec-7711-43c6-...	070231cf-7fd8-4964-acba-1b2038cbde39	5/19/2023, 2:02:55 PM	Submitting	dhsparklink	Batch job	All Attempts	22	1m 7s
Execute_206f02ec-7711-43c6-...	070231cf-7fd8-4964-acba-1b2038cbde39	5/19/2023, 1:02:53 PM	✓ Succeeded	dhsparklink	Batch job	All Attempts	21	5m 47s
Execute_35780dbf-ff3d-4843-a...	070231cf-7fd8-4964-acba-1b2038cbde39	5/19/2023, 12:58:13 PM	✓ Succeeded	dhsparklink	Batch job	All Attempts	20	5m 32s
Execute_206f02ec-7711-43c6-...	070231cf-7fd8-4964-acba-1b2038cbde39	5/19/2023, 12:03:20 PM	✓ Succeeded	dhsparklink	Batch job	All Attempts	19	6m 34s
Execute_e4791b42-7f0a-48f9-a...	070231cf-7fd8-4964-acba-1b2038cbde39	5/18/2023, 10:27:20 PM	✓ Succeeded	dhsparklink	Batch job	All Attempts	18	6m 3s
Execute_45893b79-89c1-4e17-...	070231cf-7fd8-4964-acba-1b2038cbde39	5/18/2023, 9:17:24 PM	✓ Succeeded	dhsparklink	Batch job	All Attempts	17	6m 58s
Execute_45893b79-89c1-4e17-...	070231cf-7fd8-4964-acba-1b2038cbde39	5/18/2023, 8:32:29 PM	✓ Succeeded	dhsparklink	Batch job	All Attempts	16	7m 14s
Execute_45893b79-89c1-4e17-...	070231cf-7fd8-4964-acba-1b2038cbde39	5/18/2023, 8:17:57 PM	✓ Succeeded	dhsparklink	Batch job	All Attempts	15	7m 14s

Conclusion

It's a welcome addition to the Synapse Link export options, but at the cost of needing to provision Spark Pools and also a $$$ cost too…this isn't a free export process so due care needs to be taken with it. The question is, is the cost involved worth being able to automate the export and merging of source Dynamics and Dataverse data into the Delta Lake format? I could not answer that for everyone, again I go back to **cost** vs **value**. If this provides value rather than a separate engineered process then great.

Synapse Link for Dataverse: Delta Lake Export Configuration

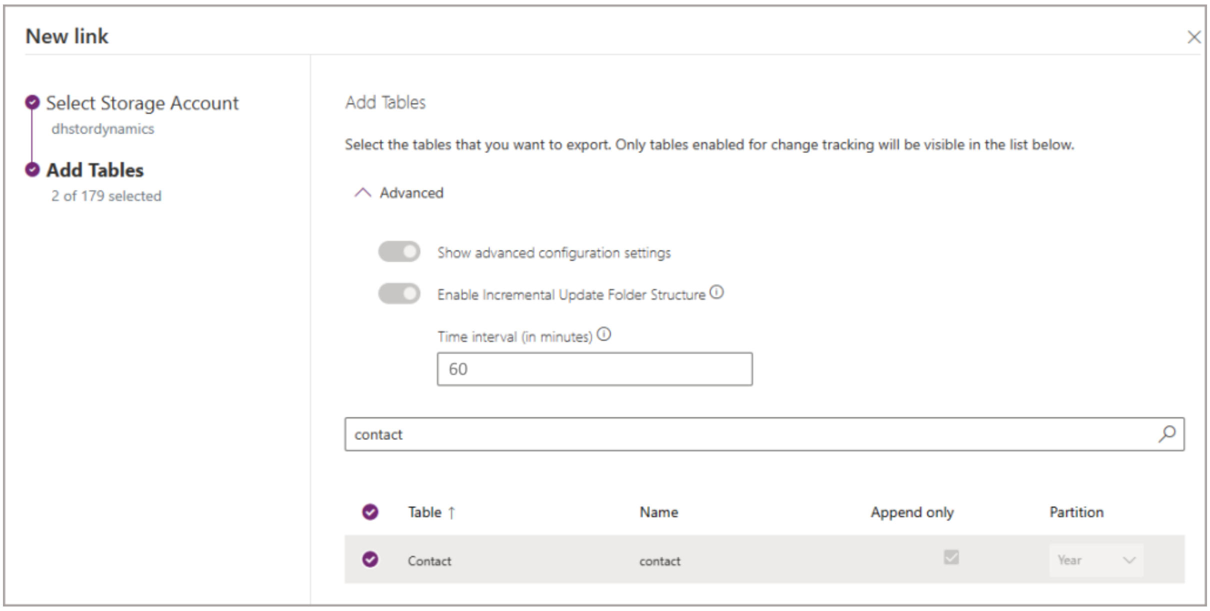
In the recently released functionality in Synapse Link for Dataverse to export to Delta Lake format, this involves creating and running Spark pools in a Synapse workspace. But when do these Spark pools run and can we control when they run? Spark pools are a chargeable service so ideally we want to be able to budget for this and work out when the pools run (and for how long).

In this section we're going to be looking at how to configure the time interval that the Spark compute will run to merge the exported Dataverse CSVs into the relevant Delta Lake folders.

Setting up Time Intervals for Delta Merge

To dictate when the Synapse Spark pools start up and write the Dataverse exported CSV data to the Delta Lake folders, use the **Time interval (in minutes)** setting in the Power Apps portal when setting up the Synapse Link and selecting tables. Please note that once this setting is configured, it can't be changed and is global across all the tables (so you can't pick and choose the time interval for each table).

If you need to change the time interval you'll need to **Unlink** the Synapse Link and start again.



New link

Select Storage Account
dhstordynamics

Add Tables
2 of 179 selected

Add Tables

Select the tables that you want to export. Only tables enabled for change tracking will be visible in the list below.

Advanced

Show advanced configuration settings

Enable Incremental Update Folder Structure

Time interval (in minutes): 60

contact

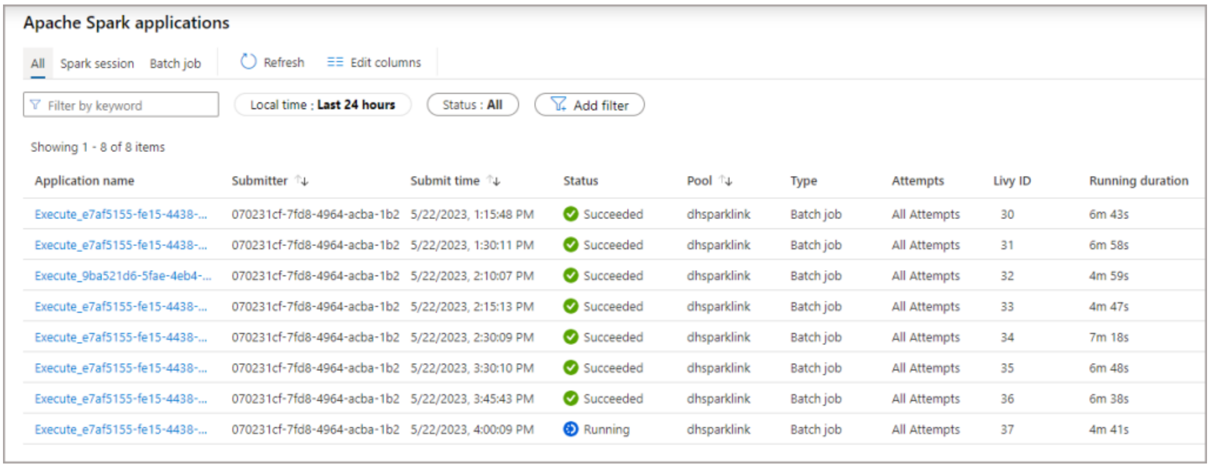
Table ↑	Name	Append only	Partition
Contact	contact	☑	Year

Monitoring Spark Applications to Check Spark Pools

Once the Synapse Link has been configured, we'll start to see Spark **batch job** activity in Synapse Studio under **Monitoring** > **Apache Spark Applications**. Use the filters to select the **Pool** that was configured in Power Apps if there are a lot of logs.

The example below shows what happens when the **Time interval** is set to **15 minutes**. If data is changed in Dynamics/Dataverse then that data with be exported to CSV and the Spark batch job will run the merge into the Delta Lake.

Please note that if no data is changed in Dynamics then the Spark jobs are **not run**, so don't worry about Spark running when it doesn't need to. You will also see a single daily batch job that runs for doing Delta maintenance, and that will happen every day regardless if any data has changed/exported from Dynamics/Dataverse.

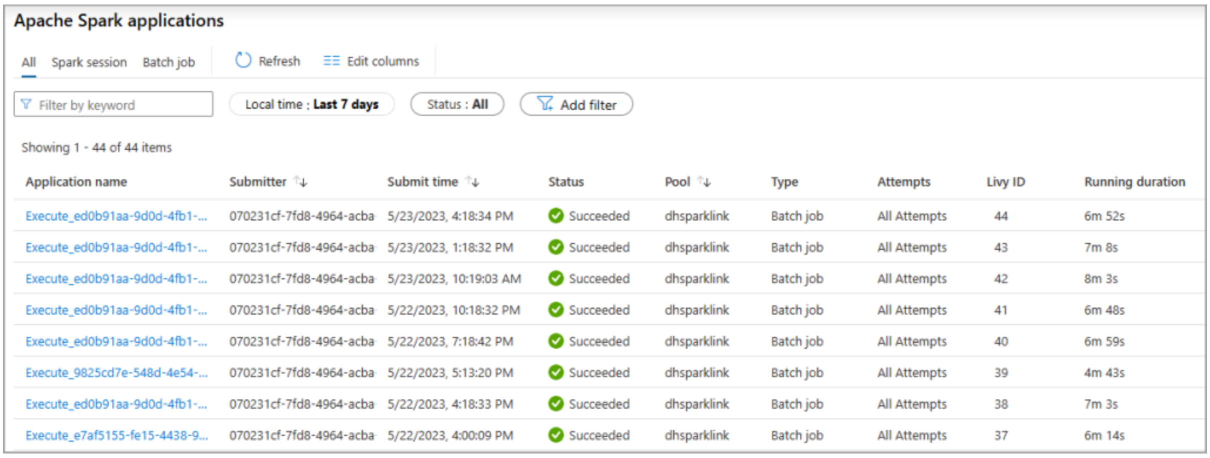


Apache Spark applications

Showing 1 - 8 of 8 items

Application name	Submitter ↑↓	Submit time ↑↓	Status	Pool ↑↓	Type	Attempts	Livy ID	Running duration
Execute_e7af5155-fe15-4438-...	070231cf-7fd8-4964-acba-1b2	5/22/2023, 1:15:48 PM	Succeeded	dhsparklink	Batch job	All Attempts	30	6m 43s
Execute_e7af5155-fe15-4438-...	070231cf-7fd8-4964-acba-1b2	5/22/2023, 1:30:11 PM	Succeeded	dhsparklink	Batch job	All Attempts	31	6m 58s
Execute_9ba521d6-5fae-4eb4-...	070231cf-7fd8-4964-acba-1b2	5/22/2023, 2:10:07 PM	Succeeded	dhsparklink	Batch job	All Attempts	32	4m 59s
Execute_e7af5155-fe15-4438-...	070231cf-7fd8-4964-acba-1b2	5/22/2023, 2:15:13 PM	Succeeded	dhsparklink	Batch job	All Attempts	33	4m 47s
Execute_e7af5155-fe15-4438-...	070231cf-7fd8-4964-acba-1b2	5/22/2023, 2:30:09 PM	Succeeded	dhsparklink	Batch job	All Attempts	34	7m 18s
Execute_e7af5155-fe15-4438-...	070231cf-7fd8-4964-acba-1b2	5/22/2023, 3:30:10 PM	Succeeded	dhsparklink	Batch job	All Attempts	35	6m 48s
Execute_e7af5155-fe15-4438-...	070231cf-7fd8-4964-acba-1b2	5/22/2023, 3:45:43 PM	Succeeded	dhsparklink	Batch job	All Attempts	36	6m 38s
Execute_e7af5155-fe15-4438-...	070231cf-7fd8-4964-acba-1b2	5/22/2023, 4:00:09 PM	Running	dhsparklink	Batch job	All Attempts	37	4m 41s

If we choose a 3 hour (180) minute time interval then when data changes are made in Dynamics/Dataverse the data is merged into the Delta Lake on that schedule. In the image below we can see 3 hours intervals between the batch jobs (there are a couple of other jobs as well as I added more tables to the sync process).

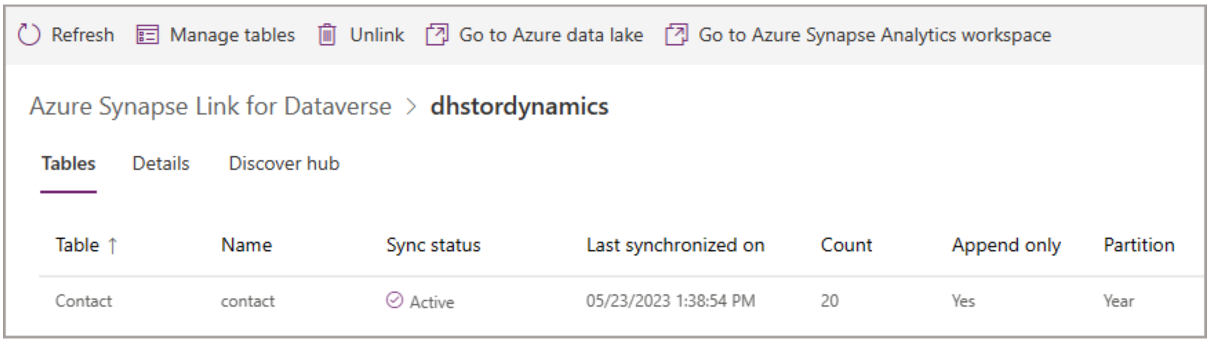


Apache Spark applications

Showing 1 - 44 of 44 items

Application name	Submitter ↑↓	Submit time ↑↓	Status	Pool ↑↓	Type	Attempts	Livy ID	Running duration
Execute_ed0b91aa-9d0d-4fb1-...	070231cf-7fd8-4964-acba	5/23/2023, 4:18:34 PM	Succeeded	dhsparklink	Batch job	All Attempts	44	6m 52s
Execute_ed0b91aa-9d0d-4fb1-...	070231cf-7fd8-4964-acba	5/23/2023, 1:18:32 PM	Succeeded	dhsparklink	Batch job	All Attempts	43	7m 8s
Execute_ed0b91aa-9d0d-4fb1-...	070231cf-7fd8-4964-acba	5/23/2023, 10:19:03 AM	Succeeded	dhsparklink	Batch job	All Attempts	42	8m 3s
Execute_ed0b91aa-9d0d-4fb1-...	070231cf-7fd8-4964-acba	5/22/2023, 10:18:32 PM	Succeeded	dhsparklink	Batch job	All Attempts	41	6m 48s
Execute_ed0b91aa-9d0d-4fb1-...	070231cf-7fd8-4964-acba	5/22/2023, 7:18:42 PM	Succeeded	dhsparklink	Batch job	All Attempts	40	6m 59s
Execute_9825cd7e-548d-4e54-...	070231cf-7fd8-4964-acba	5/22/2023, 5:13:20 PM	Succeeded	dhsparklink	Batch job	All Attempts	39	4m 43s
Execute_ed0b91aa-9d0d-4fb1-...	070231cf-7fd8-4964-acba	5/22/2023, 4:18:33 PM	Succeeded	dhsparklink	Batch job	All Attempts	38	7m 3s
Execute_e7af5155-fe15-4438-9...	070231cf-7fd8-4964-acba	5/22/2023, 4:00:09 PM	Succeeded	dhsparklink	Batch job	All Attempts	37	6m 14s

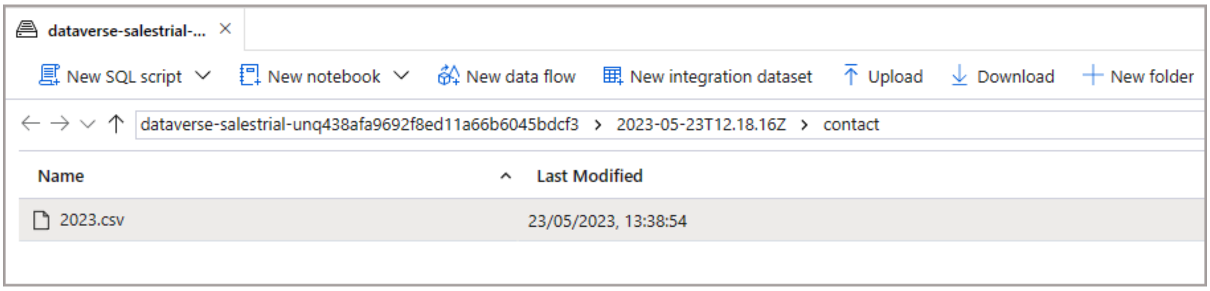
Although the Delta Lake merge process only happens as per the Synapse Link time interval, the data is actually exported to CSV in "near real-time." This CSV data isn't accessible via the Lake Database so you still need to wait until the Spark batch job is run for the CSV data to be merged into the relevant Delta Lake tables. Then you can query the relevant tables using Serverless SQL Pools or Spark pools.



Azure Synapse Link for Dataverse > **dhstordynamics**

Table ↑	Name	Sync status	Last synchronized on	Count	Append only	Partition
Contact	contact	Active	05/23/2023 1:38:54 PM	20	Yes	Year

The CSV data is exported from Dynamics/Dataverse in near real-time, but is only merged into Delta Lake as per the time interval in the Synapse Link.



dataverse-salestrial-unq438afa9692f8ed11a66b6045bdcf3 › 2023-05-23T12.18.16Z › contact

Name	Last Modified
2023.csv	23/05/2023, 13:38:54

Conclusion

In this section we’ve looked at how to configure the frequency in which the Spark pools will start and merge the exported CSV data into Delta Lake, this allows you to control when the Spark jobs run and help manage costs.