Change data capture in Azure Data Factory and Azure Synapse Analytics: -

Azure Data Factory (ADF) has recently added many new CDC-enabled connectors to process change data from SQL, Storage, Cosmos DB, and many other sources to continuously detect changes at the source.

Azure Data Factory is introducing a new mechanism to make the life of a data engineer easier. By automatically detecting data changes at the source without requiring complex designing or coding, ADF is making it a breeze to scale these processes. Change Data Capture will now exist as a **new native top-level resource** in the Azure Data Factory studio where data engineers can quickly configure continuously running jobs to process big data at scale with extreme efficiency.

The new Change Data Capture resource in ADF allows for full fidelity change data capture that continuously runs in near real-time.

Supported data sources:

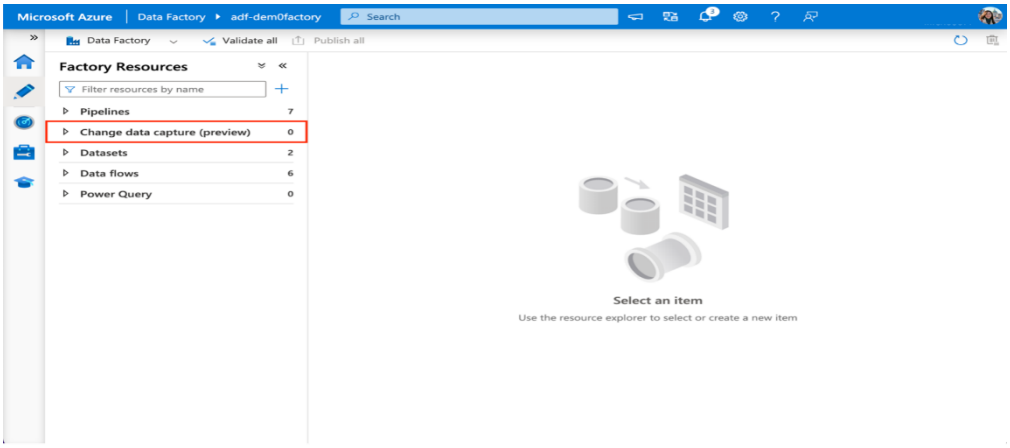
- Avro
- Azure Cosmos DB (SQL API)
- Azure SQL Database
- Delimited Text
- JSON
- ORC
- Parquet
- SQL Server
- XML

Supported targets:

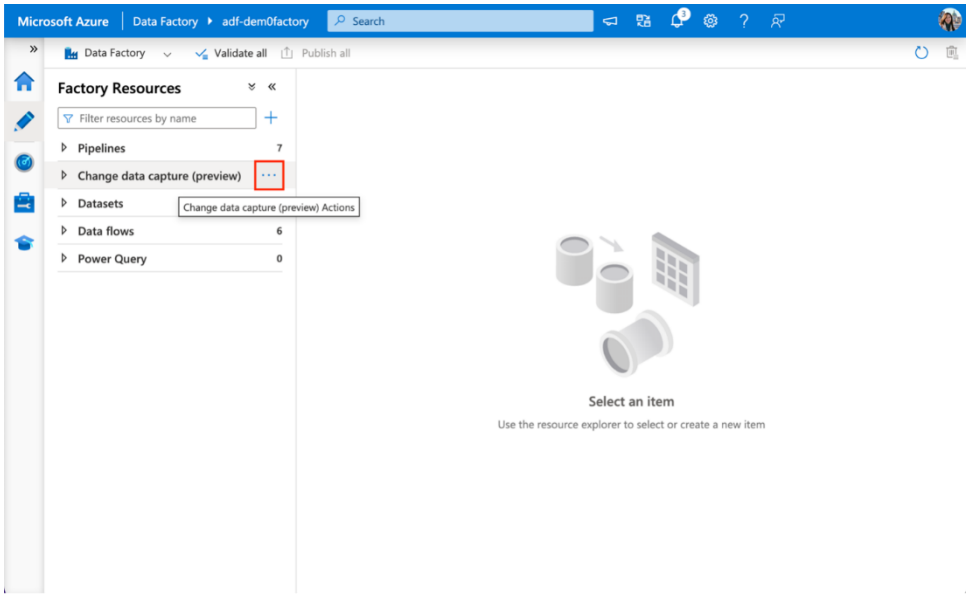
- Avro
- Azure SQL Database
- Azure Synapse Analytics
- Delimited Text
- Delta
- JSON
- ORC
- Parquet

Create a change data capture artifact:

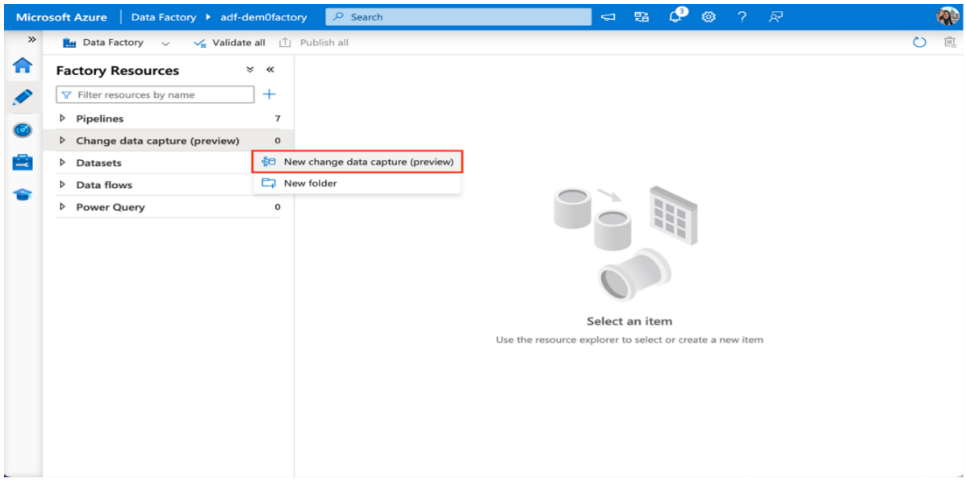
1. Navigate to the **Author** blade in your data factory. You will see a new top-level artifact under **Pipelines** called **Change data capture (preview)**.



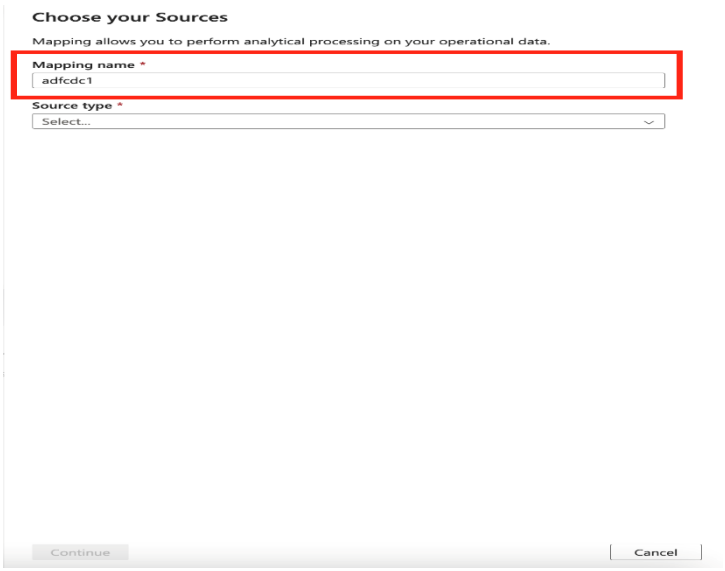
2. To create a new **Change data capture**, hover over **Change data capture (preview)** until you see 3 dots appear. Click on the **Change data capture actions**.



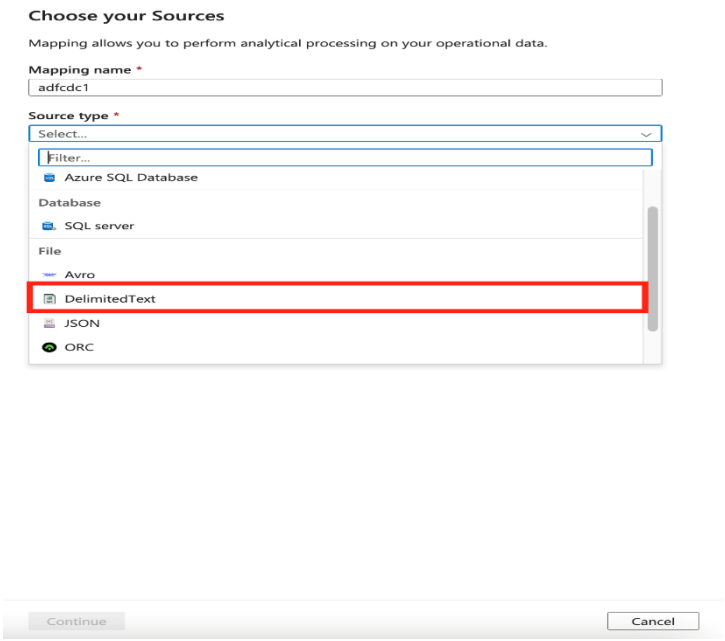
3. Select **New change data capture (preview)**. This will open a flyout to begin the guided process.



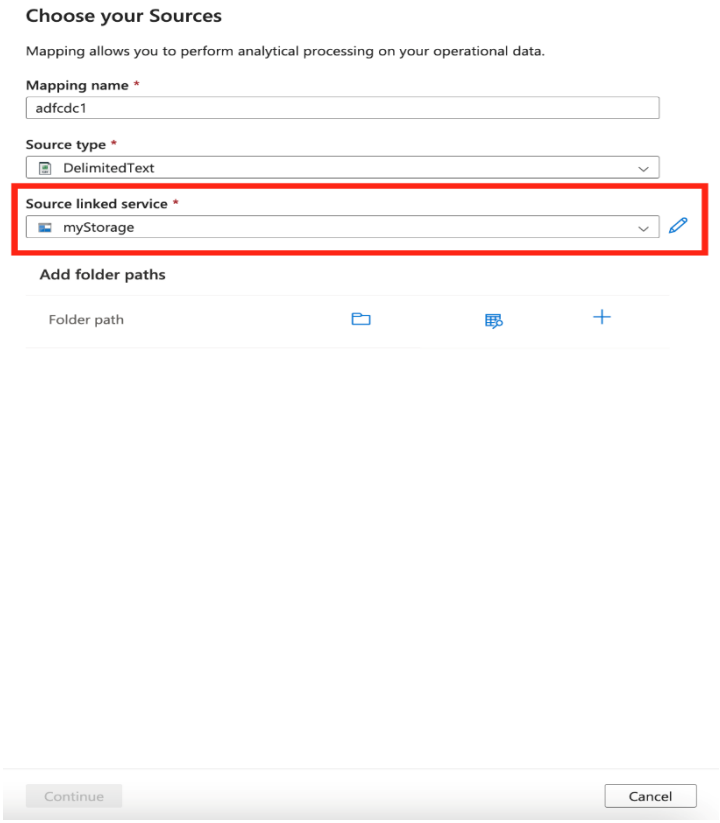
4. You will then be prompted to name your CDC resource. By default, the name will be set to “adfcdc” and continue to increment up by 1. You can replace this default name with your own.



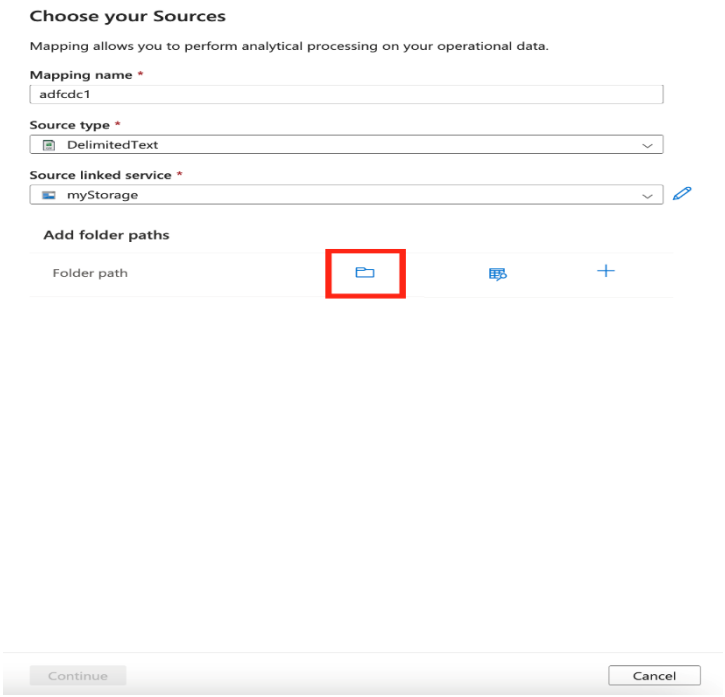
5. Use the drop-down selection list to choose your data source. For this tutorial, we will use **Delimited Text**.



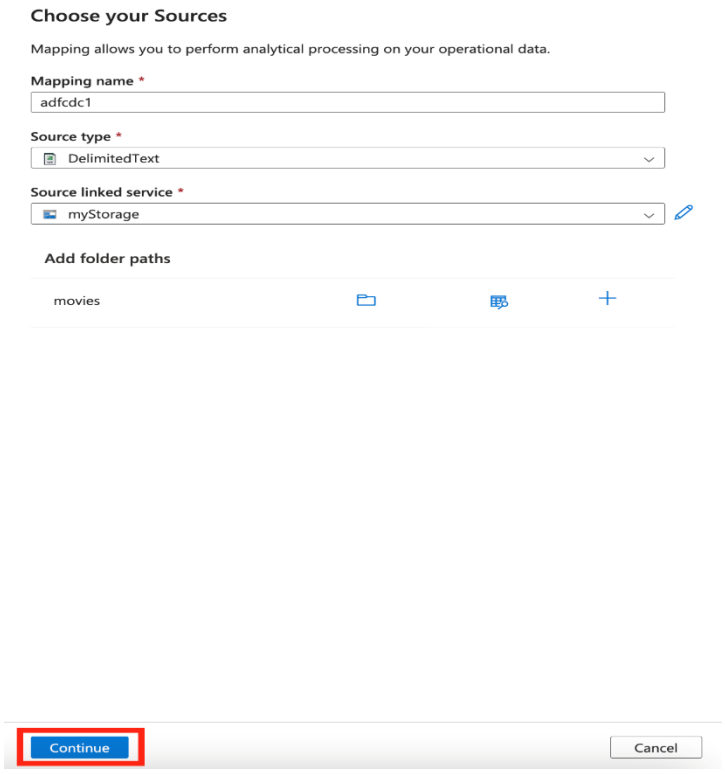
6. You will then be prompted to select a linked service. Create a new linked service or select an existing one.



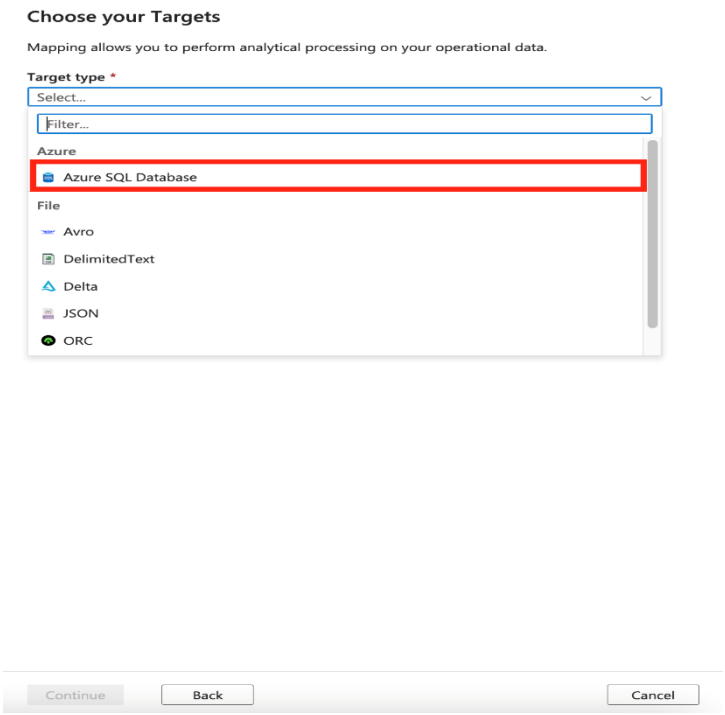
7. Use the **Browse** button to select your source data folder.



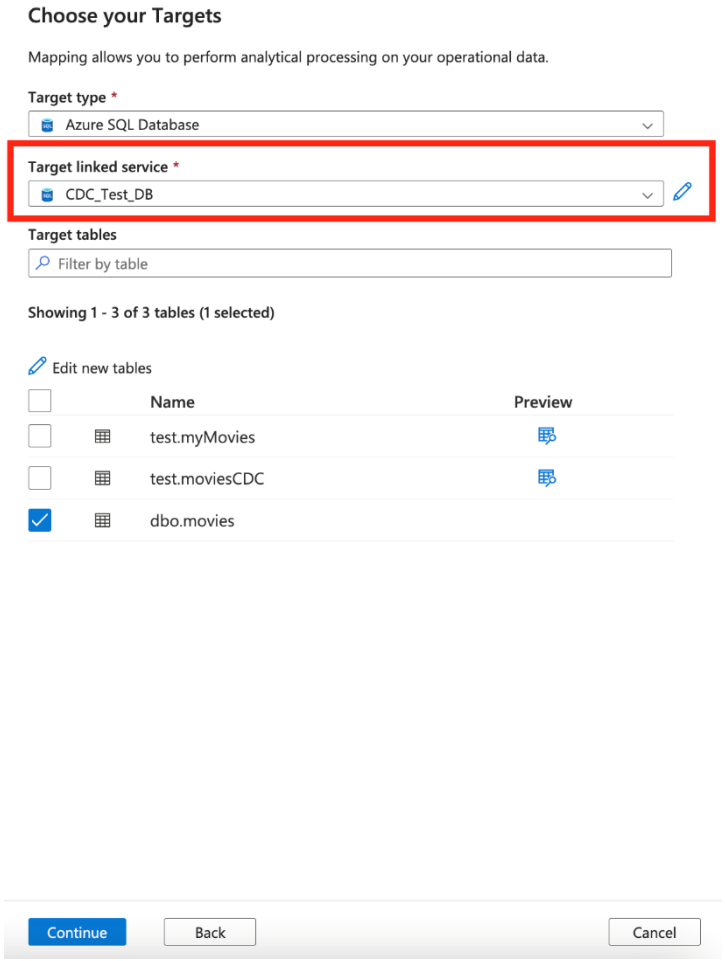
8. Once you've selected a folder path, click **Continue** to set your data target.



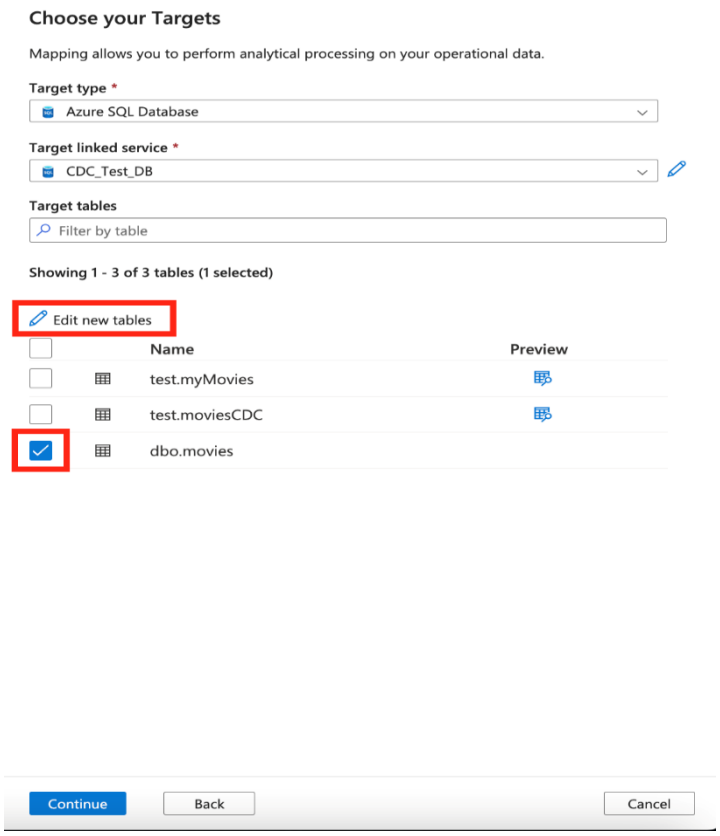
9. Then, select a **Target type** using the drop-down selection. For this tutorial, we will select **Azure SQL Database**.



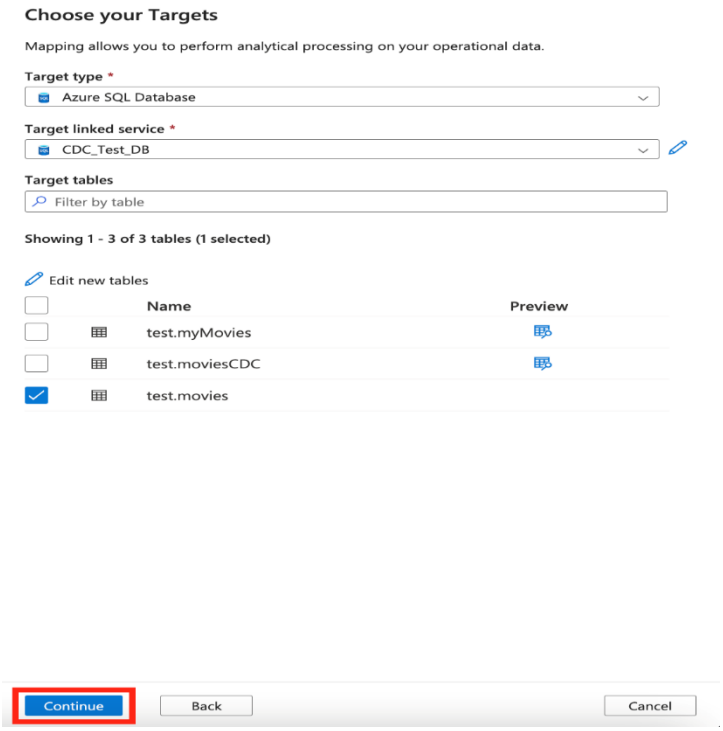
10. You will then be prompted to select a linked service. Create a new linked service or select an existing one.



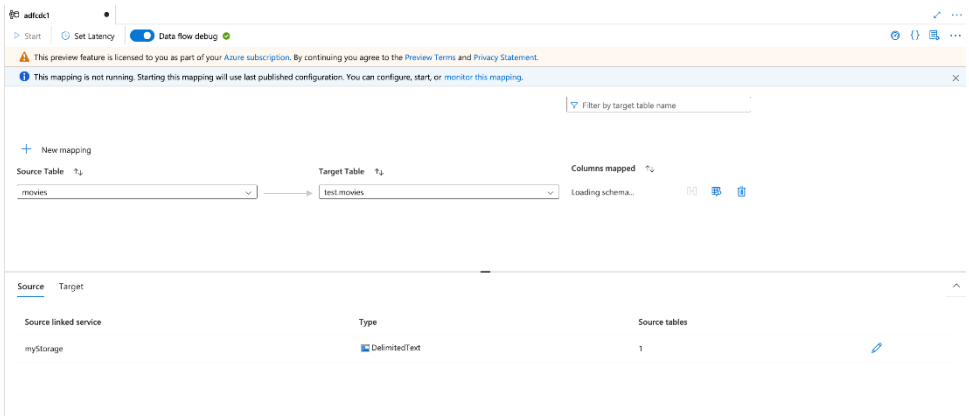
11. Create new **Target table(s)** or select an existing **Target table(s)**. Use the checkbox to make your selection(s). The **Preview** button will allow you to view your table data.



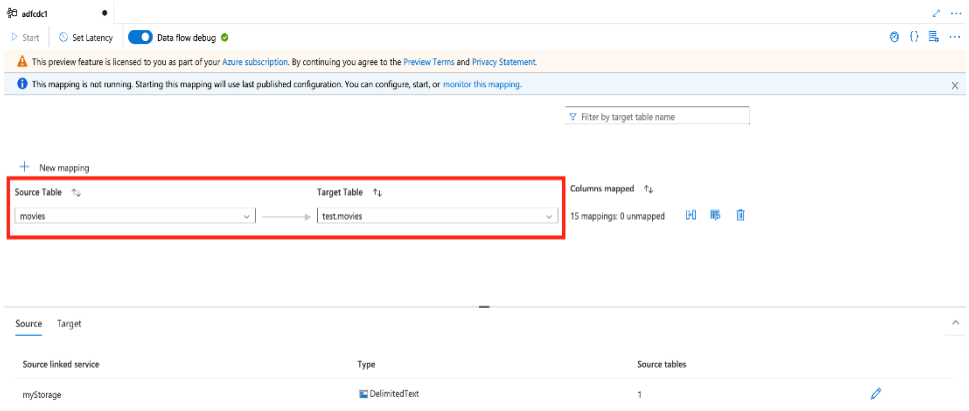
12. Click **Continue** when you have finalized your selection(s).



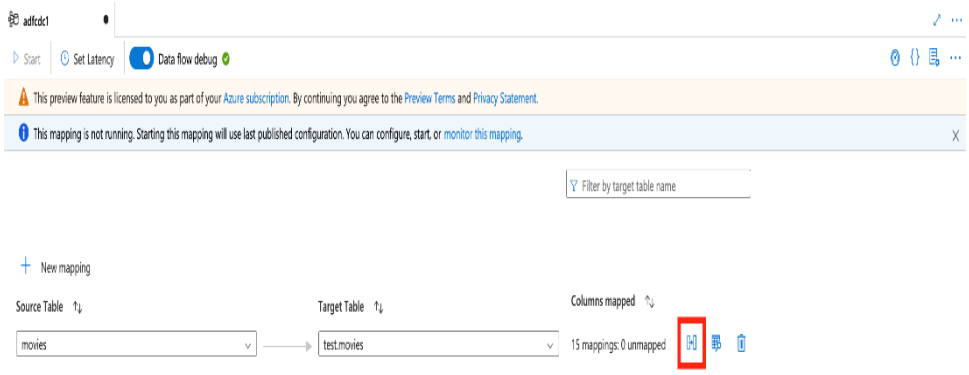
13. You will automatically land in a new change data capture tab, where you can configure your new resource.



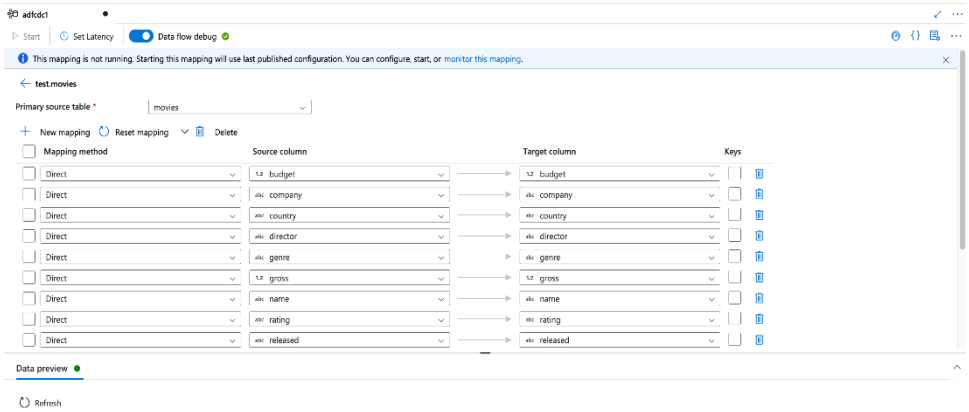
14. A new mapping will automatically be created for you. You can update the **Source** and **Target** selections for your mapping by using the drop-down selection lists.



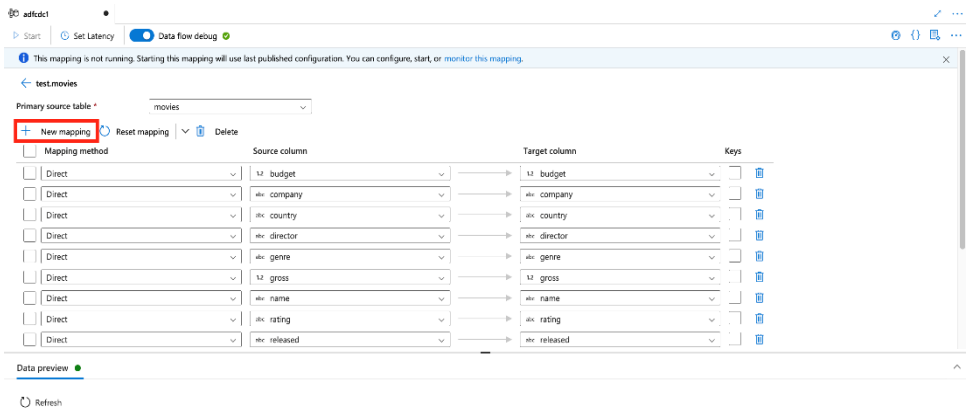
15. Once you've selected your tables; you should see that there are columns mapped. Select the **Column mappings** button to view the column mappings.



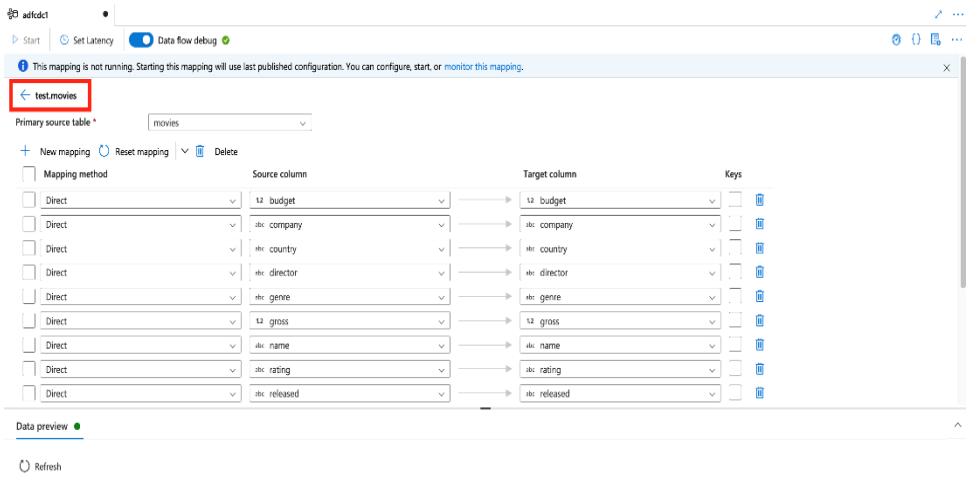
16. Here you can view your column mappings. Use the drop-down lists to edit your column mappings for **Mapping method**, **Source column**, and **Target** column.

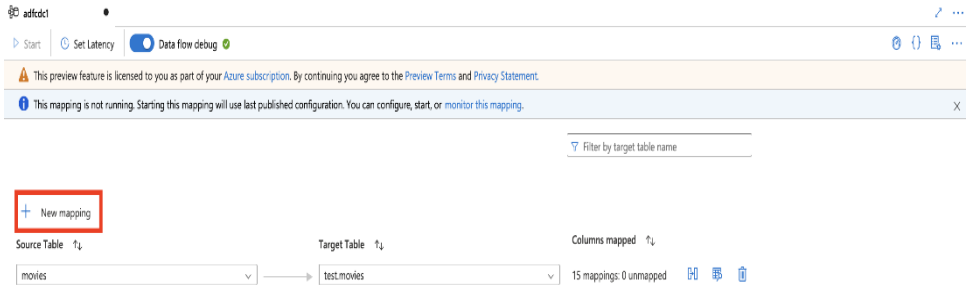


17. You can add additional column mappings using the **new mapping** button. Use the drop-down lists to select the **Mapping method**, **Source column**, and **Target** column.

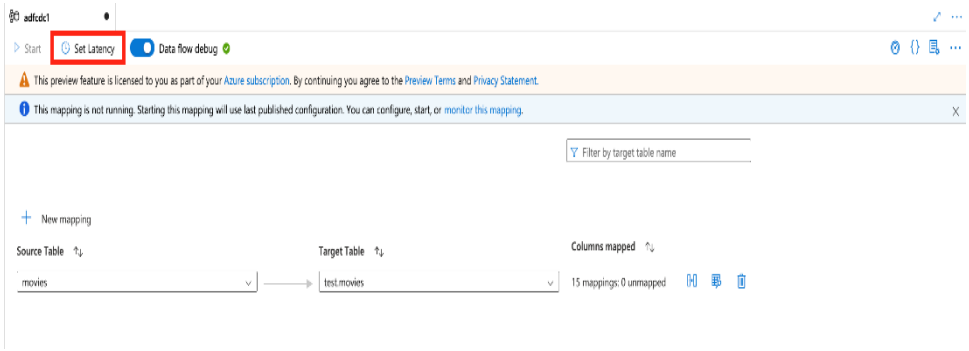


18. When your mapping is complete, click the back arrow to return to the main canvas.



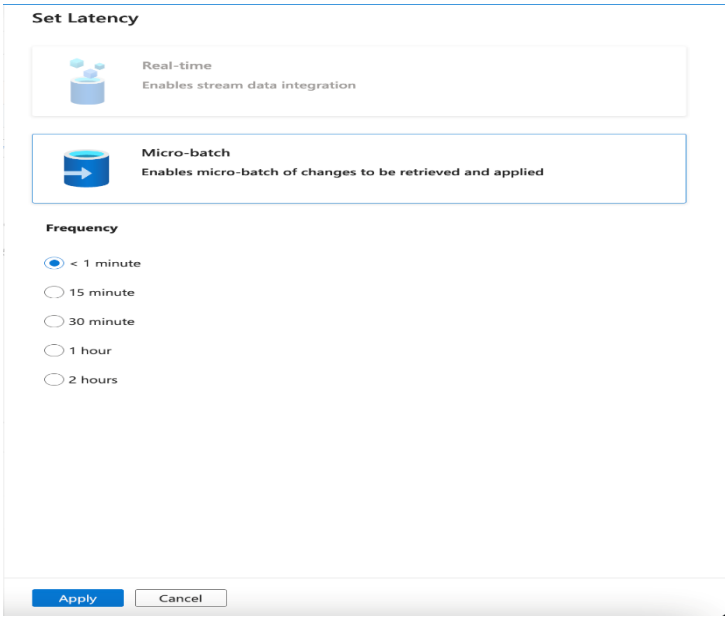


19. Once your mapping complete, set your frequency using the **Set Latency** button.

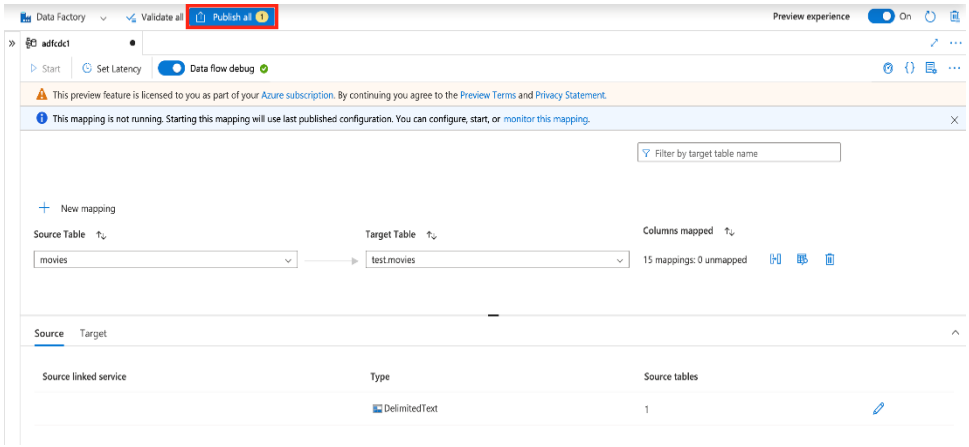


20. Select the cadence of your change data capture and click **Apply** to make the changes. By default, it will be set to 15 minutes.

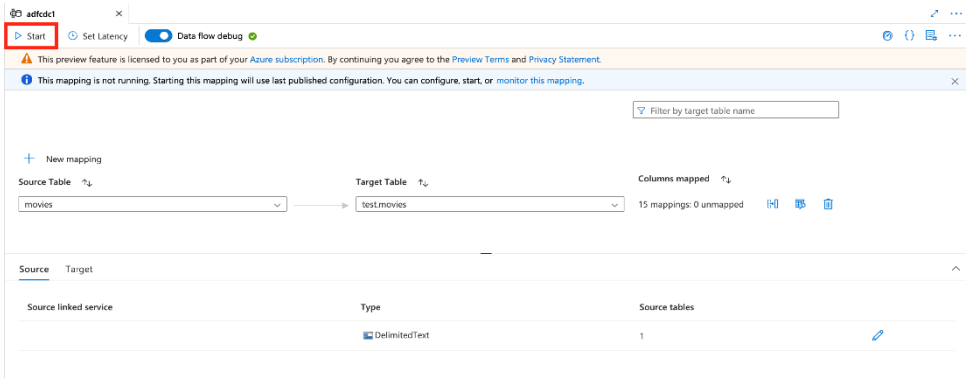
For example, if you select 30 minutes, every 30 minutes, your change data capture will process your source data and pick up any changed data since the last processed time.

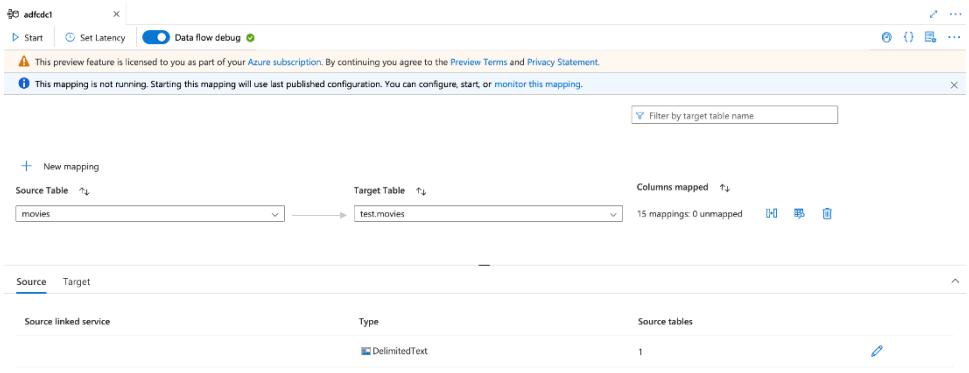


21. Once everything has been finalized, publish your changes.



22. Click **Start** to start running your **Change data capture**.





23. To monitor your change data capture, navigate to the **Monitor** blade or click the monitoring icon from the CDC designer.

Scenario 1(Source Data and Sink Data before CDC): -

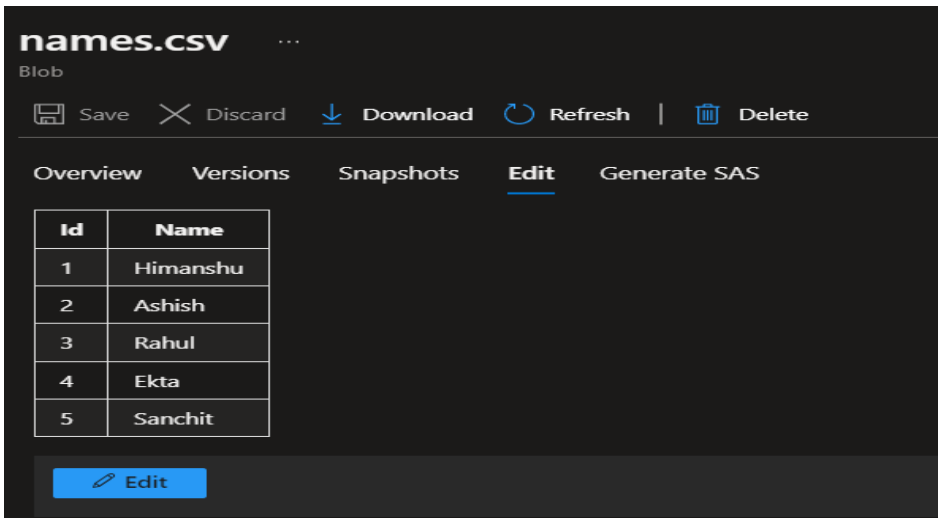


Fig: Source Data CSV File (Blob Storage)

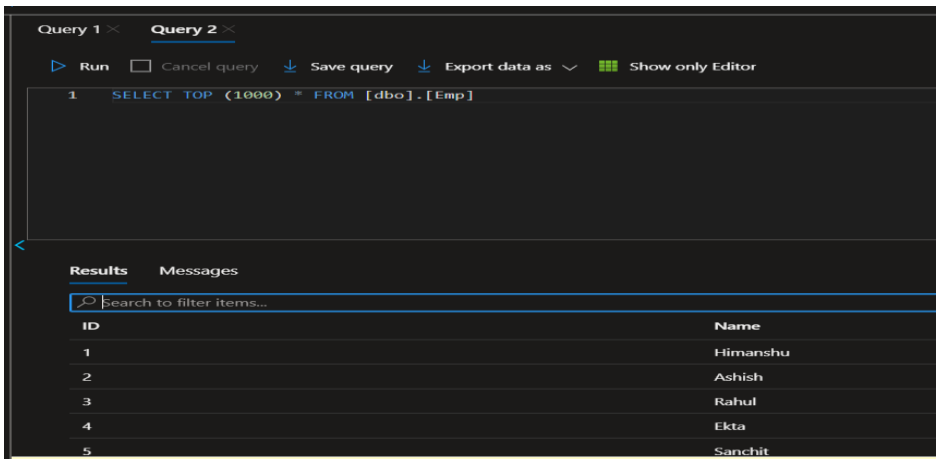


Fig: Sink Data Table (Azure SQL DB)

Scenario 2(Source Data and Sink Data After CDC Template Execution):

We have added 2 new rows in Source CSV File (Blob Storage) highlighted with Yellow Colour. Only these 2 rows should come in the execution Of CDC Template to meet the set criteria of CDC.

We have set up a latency window of 15 minutes so every 15 minutes it will check for new records in the Source File.

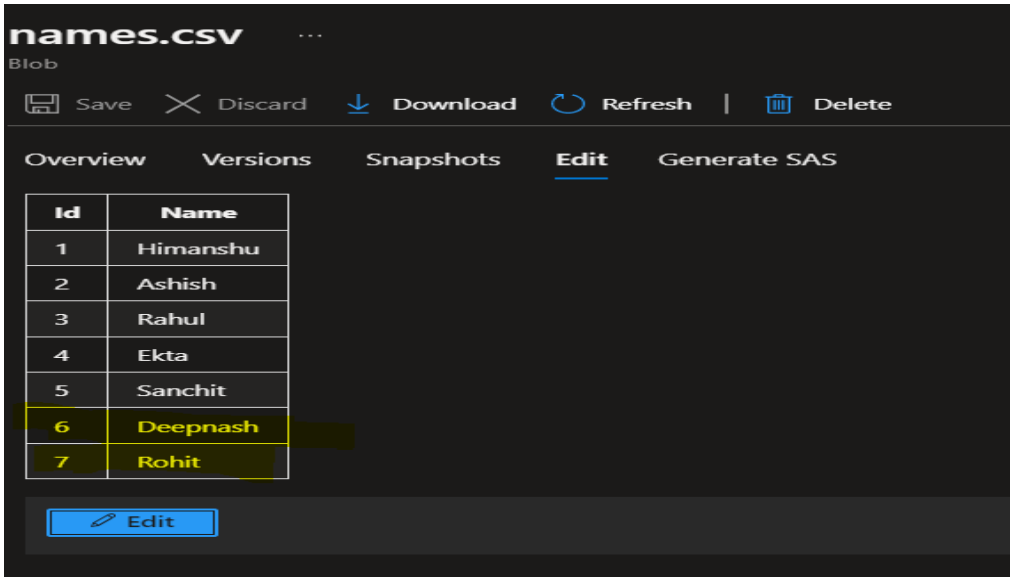


Fig: Updated Source Data

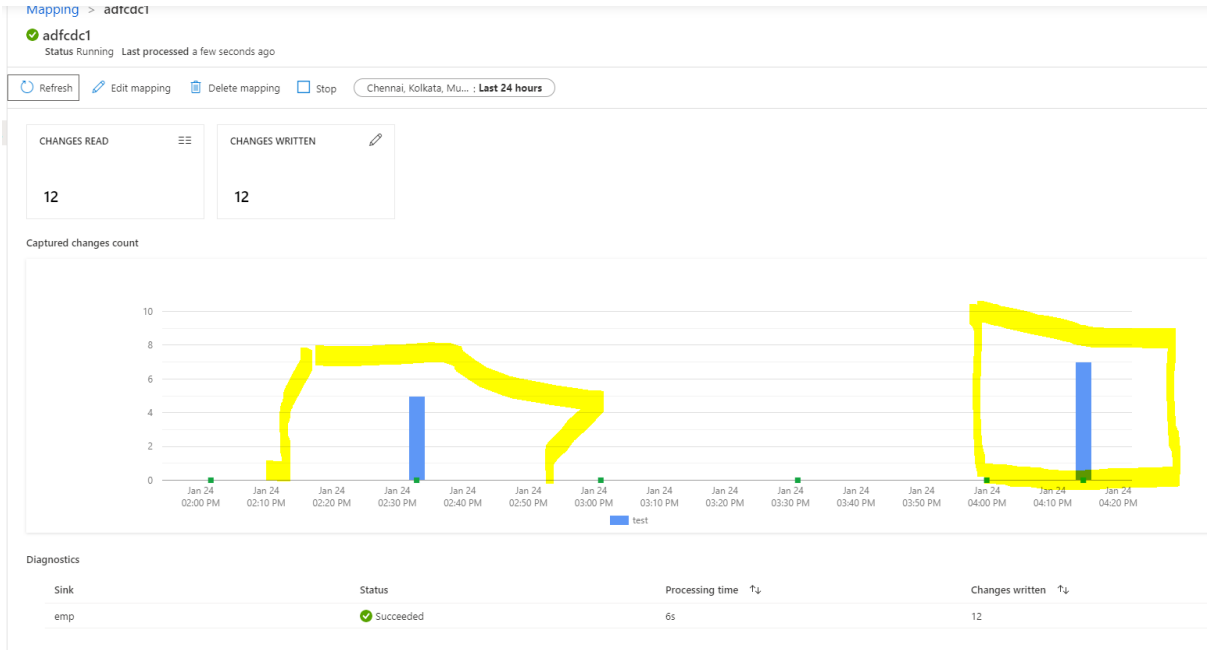


Fig: CDC Captured at Source Side

Limitations:

- Currently, when creating source/target mappings, each source and target is only allowed to be used once.
- Continuous, real-time streaming is coming soon.
- Allow schema drift is coming soon.

Advantages:

- Change Data Capture resource in ADF allows for full fidelity change data capture that continuously runs in near real-time through.
- Data engineers can quickly configure continuously running jobs to process big data at scale with extreme efficiency.
- With the CDC resource, you will not bead to design pipelines or data flow activities and the only billing will be 4 cores of General-Purpose data flows while your data in being processed. hat is the only time you will be billed.