

Case Study 4 – Dynamic Ingestion of Tables using Metadata Control in Azure Data Factory

by Sudarshan Zunja

Setting Configurations and Environments

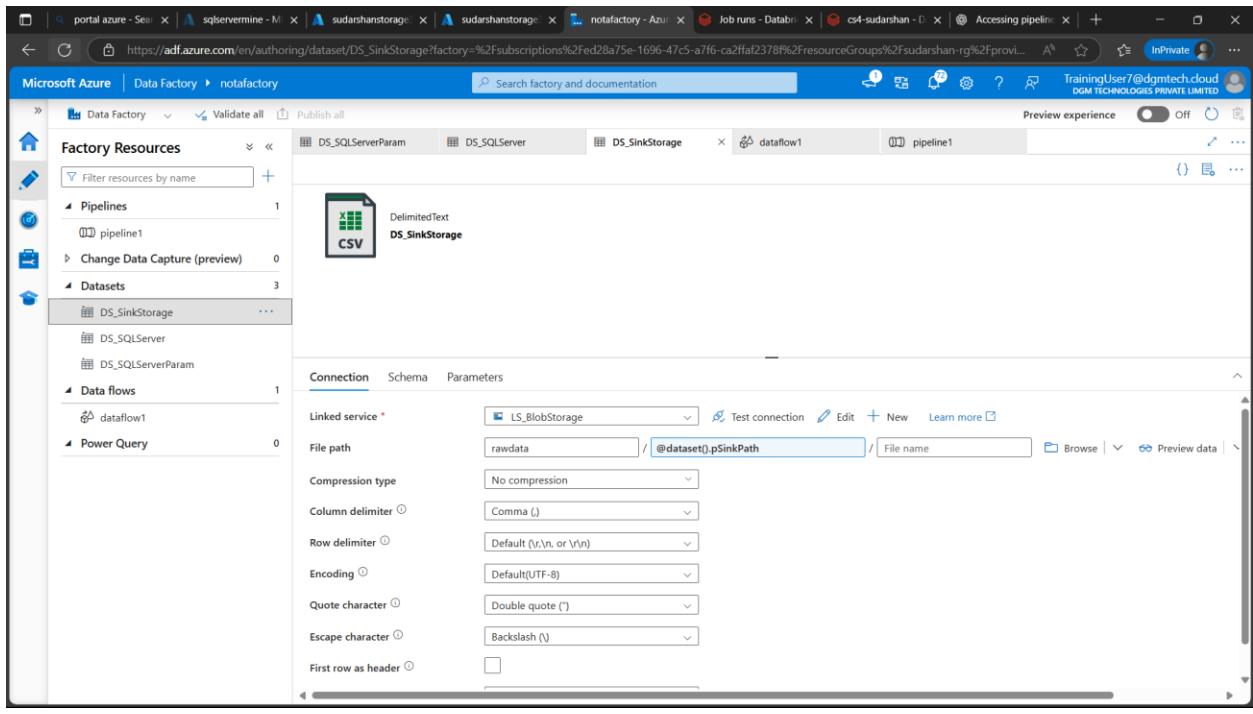
Source Dataset – SQL Server Database Tables

Target Sink – Azure Blob Storage Container

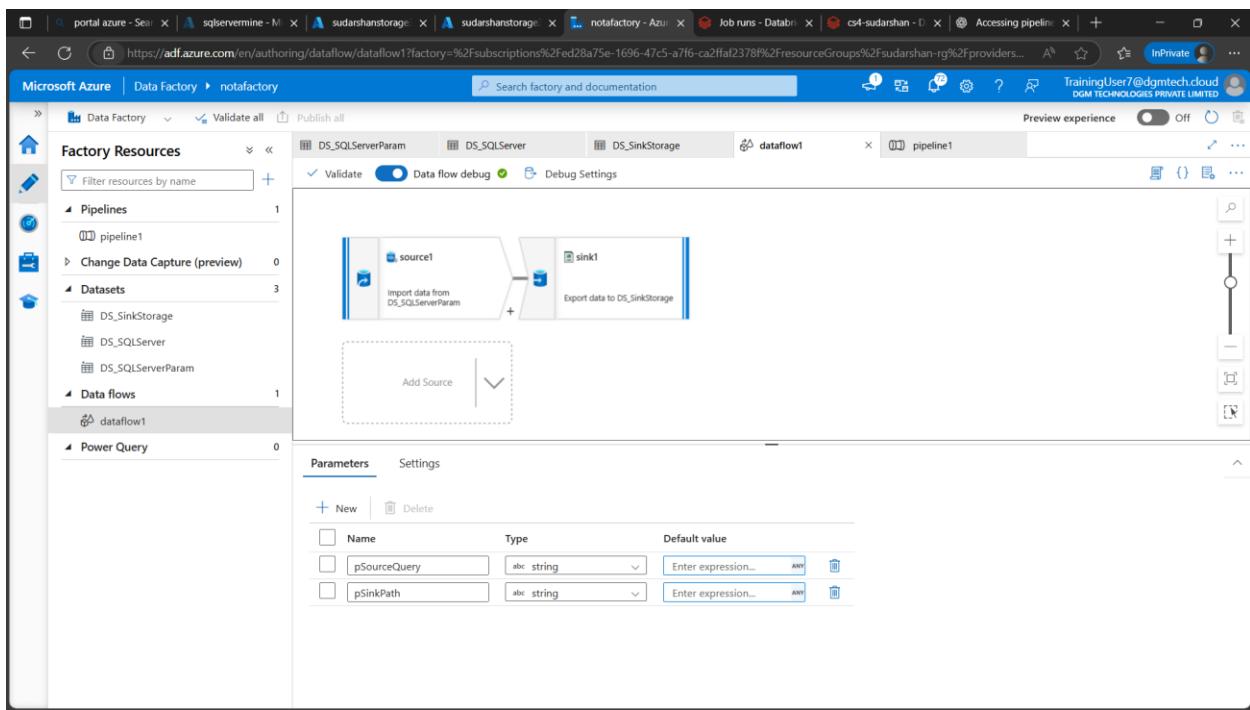
Transformation Intermediaries – Azure Databricks

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under Datasets, 'DS_SQLServerParam' is selected. The main pane displays the configuration for 'DS_SQLServerParam'. It includes tabs for Connection, Schema, and Parameters. Under Connection, the 'Linked service' dropdown is set to 'LS_SQLServerParam' and the 'Table' dropdown is set to 'dbo/metadatacontrol2'. There is also a checkbox for 'Enter manually'. The top navigation bar shows the URL https://adf.azure.com/en/authoring/dataset/DS_SQLServerParam?factory=%2Fsubscriptions%2Fed28a75e-1696-47c5-a7f6-ca2ffa2378f%2FresourceGroups%2Fsudarshan-rg%2F....

This screenshot is nearly identical to the one above, showing the same Microsoft Azure Data Factory interface. The 'Factory Resources' sidebar and the configuration for 'DS_SQLServerParam' are identical. The main pane shows the 'Connection' tab with 'Linked service' set to 'LS_SQLServerParam' and 'Table' set to 'dbo/metadatacontrol2'. The top navigation bar shows the URL https://adf.azure.com/en/authoring/dataset/DS_SQLServerParam?factory=%2Fsubscriptions%2Fed28a75e-1696-47c5-a7f6-ca2ffa2378f%2FresourceGroups%2Fsudarshan-rg%2F....



Data Flow Structure

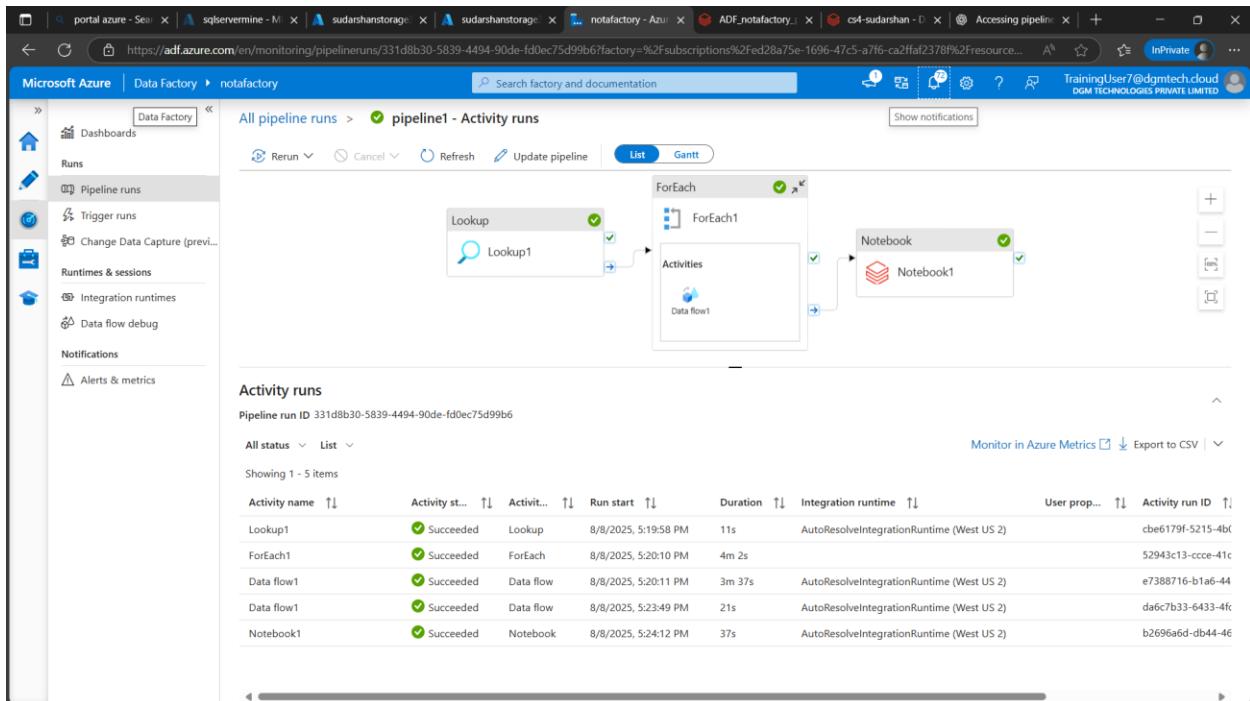


The screenshot shows the Microsoft Azure Data Factory Data Flow designer interface, focusing on the configuration of the 'source1' stage. The 'Source settings' tab is active. The 'Output stream name' is set to 'source1'. The 'Description' field contains the text 'Import data from DS_SQLServerParam'. Under 'Source type', the 'Dataset' option is selected. The 'Dataset' dropdown is set to 'DS_SQLServerParam'. In the 'Options' section, the 'Allow schema drift' checkbox is checked. The 'Sampling' section has the 'Disable' radio button selected. Other tabs like 'Source options', 'Projection', 'Optimize', 'Inspect', and 'Data preview' are also visible.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (pipeline1), 'Datasets' (DS_SinkStorage, DS_SQLServer, DS_SQLServerParam), 'Data flows' (dataflow1), and 'Power Query' (0). The main workspace displays a data flow named 'dataflow1'. The flow starts with a 'source1' component (Import data from DS_SQLServerParam) and ends at a 'sink1' component (DS_SinkStorage). The sink has a description 'Export data to DS_SinkStorage'. The pipeline is currently in 'Data flow debug' mode, indicated by a green gear icon. The 'Sink' tab is selected, showing configuration options like 'Output stream name' (sink1), 'Description' (Export data to DS_SinkStorage), 'Incoming stream' (source1), 'Sink type' (Dataset, Inline, Cache), 'Dataset' (DS_SinkStorage), and 'Options' (Allow schema drift checked). A 'Validate schema' checkbox is also present.

The screenshot shows the Microsoft Azure Data Factory interface under the 'notafactory' factory. The left sidebar includes sections for General, Factory settings, Connector upgrade advis..., Connections, Linked services, Source control, Author, Security, Credentials, Customer managed key, Outbound rules, and Managed private endpoints. The 'Connections' section is currently selected. The main area displays the 'Linked services' section, which defines connection information to data stores or compute. It shows four items: 'AzureDatabricks1' (Type: Azure Databricks, Related: 1), 'AzureDatabricksDeltaLake1' (Type: Azure Databricks Delta Lake, Related: 0), 'LS_BlobStorage' (Type: Azure Data Lake Storage Gen2, Related: 2), and 'LS_SQLServerParam' (Type: SQL Server, Related: 2). A 'New' button is available for creating new linked services.

Pipeline Structure and Run



Databricks Transformations and Jobs Runs

Microsoft Azure **databricks**

Search data, notebooks, recents, and more... CTRL + P

sudarshan-databricks

Jobs & Pipelines

Create new

- Ingestion pipeline**: Ingest data from popular apps, databases and file sources
- ETL pipeline**: Build ETL pipelines using SQL and Python
- Job**: Orchestrate notebooks, pipelines, queries and more

Jobs & pipelines **Job runs** Send feedback

Job Run as Start: 08/06/2025 05:30 PM End: 08/08/2025 05:30 PM Run status Error code

Top 5 error codes (2 errors)
RunExecutionError 2

Start time	Job	Run as	Launched	Duration	Status	Error code	Run parameters
Aug 08, 2025, 05:24 PM	ADF_notafactory_pipeline1	TrainingUser7	By runs submit API	16s	Succeeded		
Aug 08, 2025, 05:17 PM	ADF_notafactory_pipeline1	TrainingUser7	By runs submit API	12s	Succeeded		
Aug 08, 2025, 05:12 PM	ADF_notafactory_pipeline1	TrainingUser7	By runs submit API	15s	Succeeded		
Aug 08, 2025, 05:10 PM	ADF_notafactory_pipeline1	TrainingUser7	By runs submit API	19s	Failed	RunExecutionError	

Create job Previous Next

Microsoft Azure **databricks**

Search data, notebooks, recents, and more... CTRL + P

sudarshan-databricks

Runs > **ADF_notafactory_pipeline1_Notebook1_b2696a6d-db44-4636-8083-cc22c3eb92a2 run** Lakeflow Jobs UI: OFF

Output Hide code Export as HTML

```
return df

pii_columns = [
    1: "hash",
    2: "hash",
    3: "redact",
    4: "redact"
]

masked_customers_df = mask_pii(cust_silver, pii_columns)
masked_customers_df.show(5)

masked_customers_df.write.mode("overwrite").format("delta").save(gold_path)

(2) Spark Jobs
+---+---+---+---+
| _c0| _c1| _c2| _c3| _c4|
+---+---+---+---+
[P1001]64604487c74ef2fb0...|55898449e674fb2e...| [REDACTED] |REDACTED|
[P1001]a4906a549b8416b9...|[0286249762f7c9434...| [REDACTED] |REDACTED|
[P1002]fdcb58571fc93e8482...|[3a7869538b838b5cc...| [REDACTED] |REDACTED|
+---+---+---+---+
```

Task run

Details Metrics

Details

- Job ID: 1049332639346510
- Task run ID: 872422149560001
- Run as: TrainingUser7
- Started: Aug 08, 2025, 05:24 PM
- Ended: Aug 08, 2025, 05:24 PM
- Duration: 16s
- Queue duration: -
- Status: Succeeded

View run events View run libraries

Notebook /Users/traininguser7@dgmtech.cloud/cs4-sudarshan

Comments

Final Medallion Layers (Bronze-Silver-Gold)

Microsoft Azure

Home > sudarshanstorage3

sudarshanstorage3 | Storage browser

Storage account

Search

Overview

Activity log

Tags

Diagnose and solve problems

Access Control (IAM)

Data migration

Events

Storage browser

Partner solutions

Resource visualizer

Data storage

Security + networking

Networking

Access keys

Shared access signature

Encryption

Microsoft Defender for Cloud

Add or remove favorites by pressing **Ctrl+Shift+F**

Enhance the security of this storage account Does this storage account follow security best practices +1

+ Add Directory Upload Refresh Delete Copy Paste Rename Acquire lease Break lease ...

Blob containers > rawdata > rawdata

Authentication method: Access key (Switch to Microsoft Entra user account)

Search blobs by prefix (case-sensitive) Only show active objects

Showing all 3 items

Name	Last modified	Access tier	Blob type	Size	Lease state
[...]	8/8/2025, 5:24:26 PM	Gold			
Gold	8/8/2025, 5:24:24 PM	Silver			
Silver	8/8/2025, 5:23:31 PM	bronzelayern			

View all

File shares

Queues

Tables

Add Directory Upload Refresh Delete Copy Paste Rename Acquire lease Break lease ...

...

...

...

...

...

...