Custom Learnings

Day 8

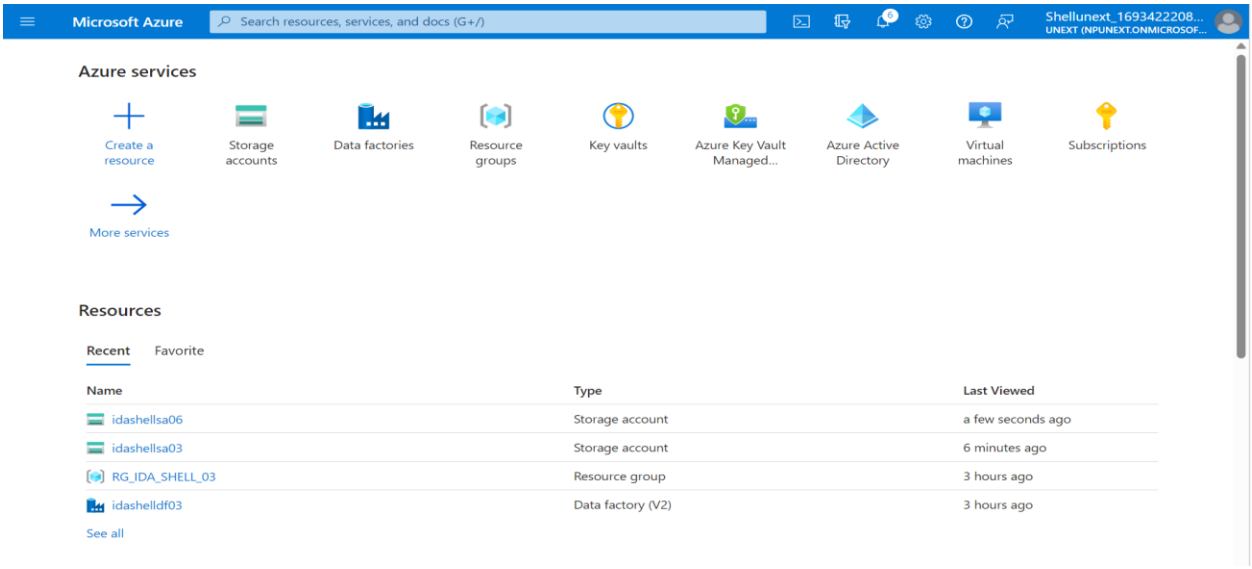
Azure Data Factory

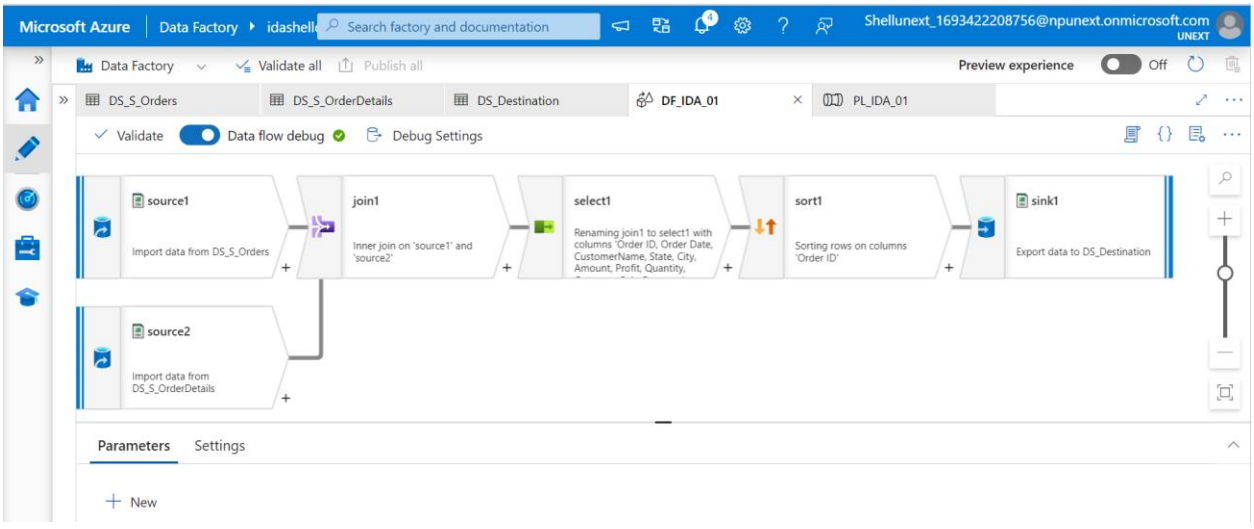
Data Flow: Used for data transformation.

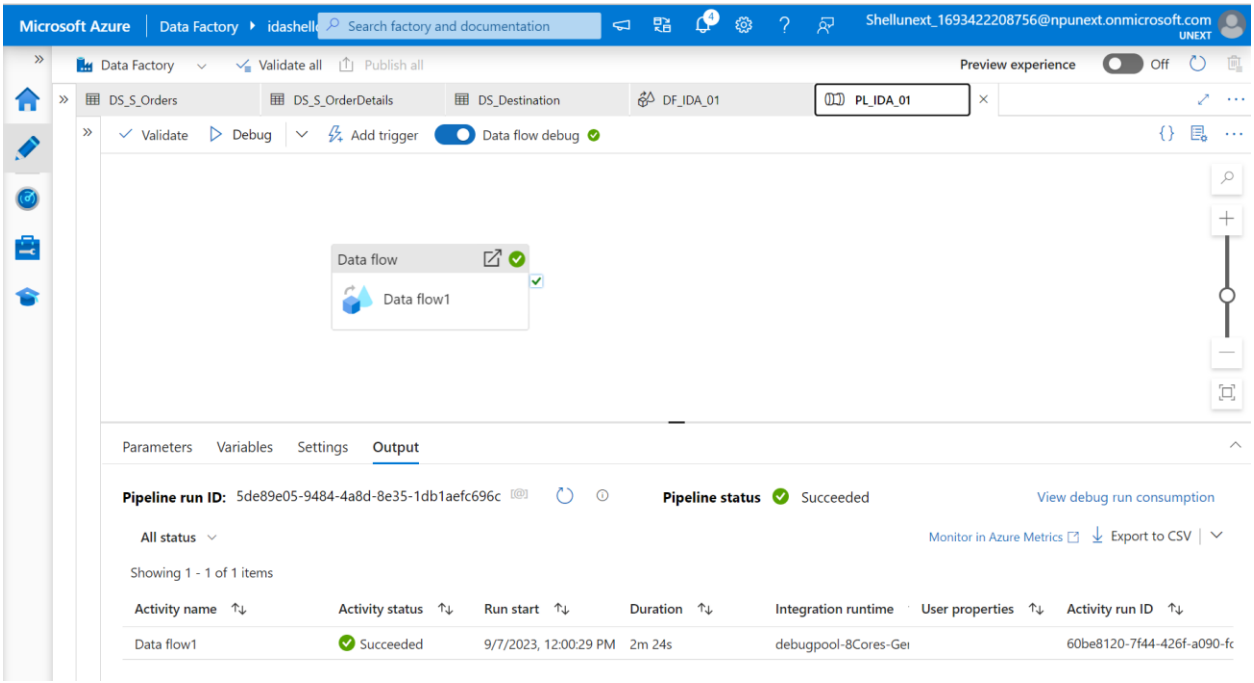
To do the transformation in Azure Data Factory it needs the cluster in the backend for spark.

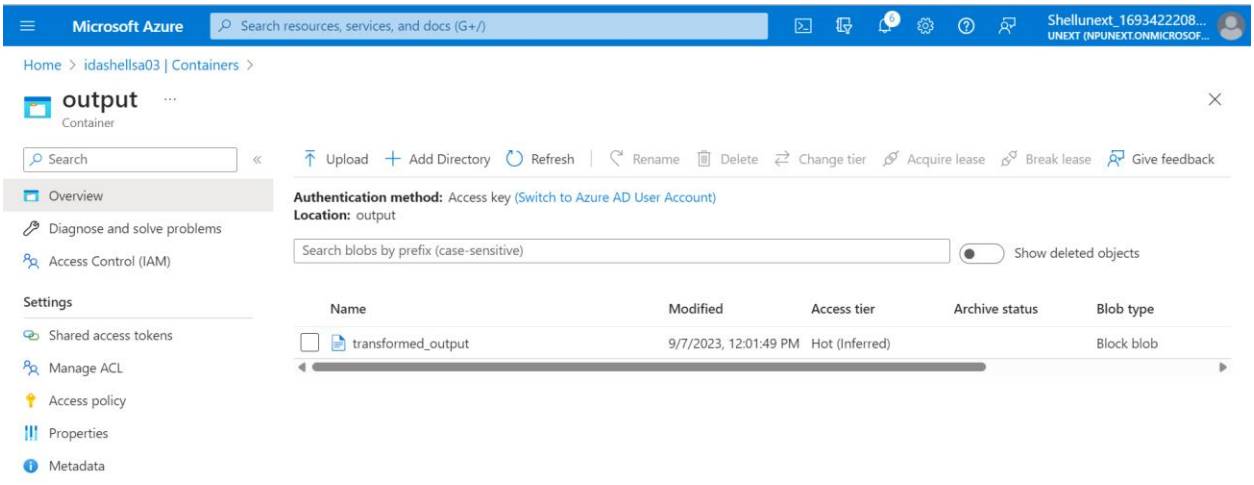
For performing some basic transformations on the data, we can use ADF but for complex and advanced transformations we use Databricks.

Creation of dataflow for taking the input data from Azure Blob Storage and storing it in ADLS after performing the transformations:

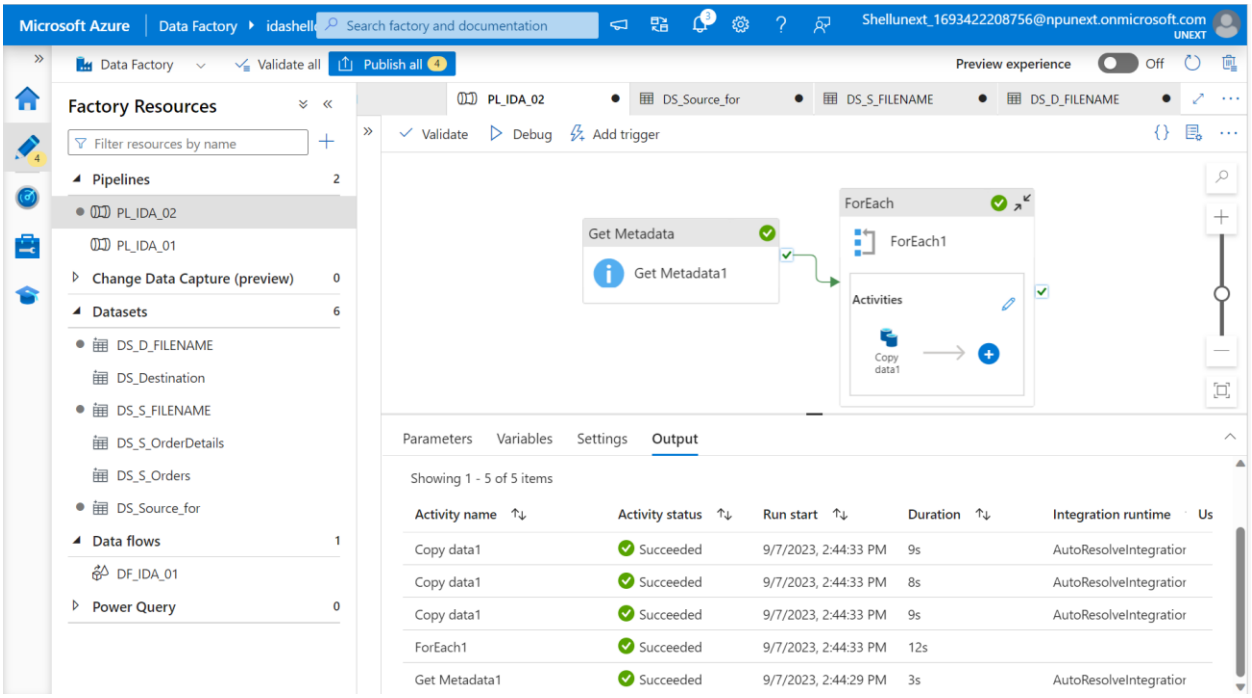


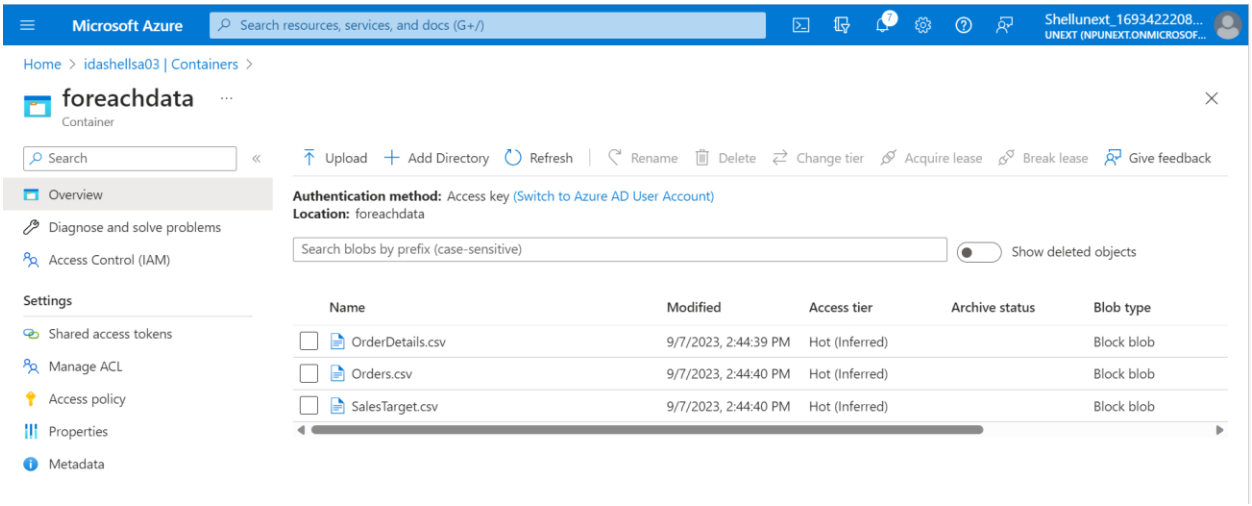




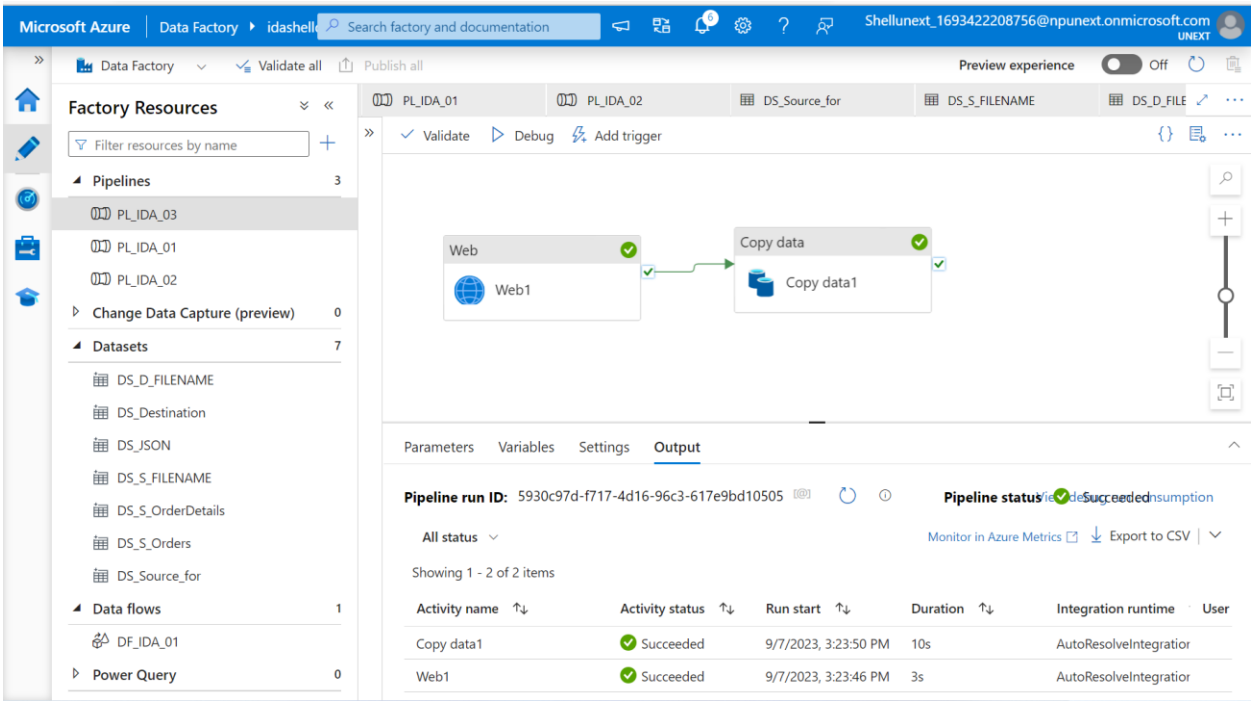


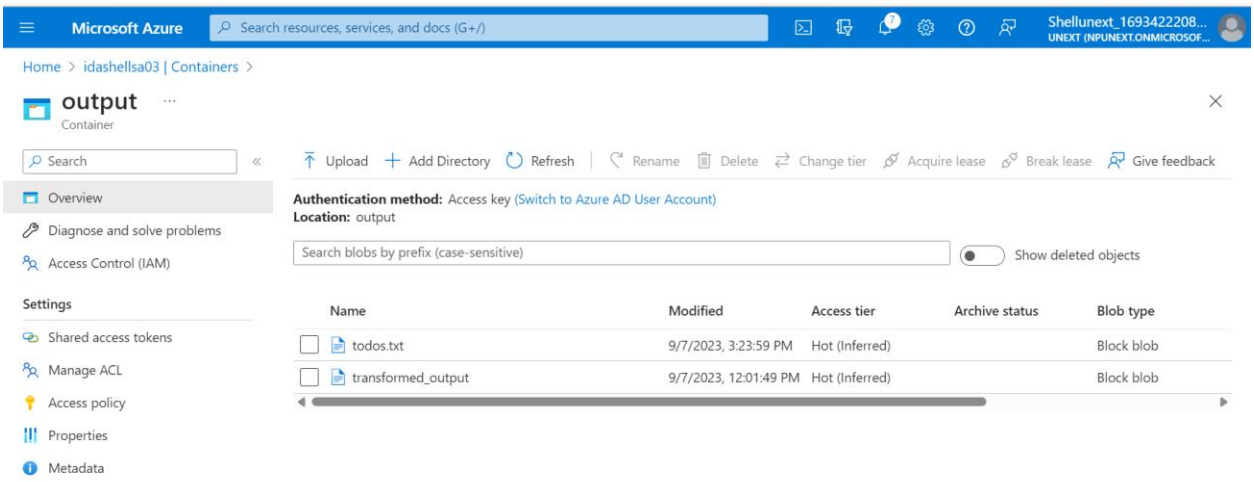
Task: to get the names of the files inside sales container and perform for each activity to copy data of each file into ADLS container.


Microsoft Azure
Data Factory
Search factory and documentation
Shellunext_1693422208756@npunext.onmicrosoft.com
UNEXT
Data Factory
Validate all
Publish all 4
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 2
PL_IDA_02
PL_IDA_01
Change Data Capture (preview) 0
Datasets 6
DS_D_FILENAME
DS_Destination
DS_S_FILENAME
DS_S_OrderDetails
DS_S_Orders
DS_Source_for
Data flows 1
DF_IDA_01
Power Query 0
PL_IDA_02
DS_Source_for
DS_S_FILENAME
DS_D_FILENAME
Validate
Debug
Add trigger
Get Metadata
Get Metadata1
ForEach
ForEach1
Activities
Copy data1
Parameters
Variables
Settings
Output
Showing 1 - 5 of 5 items
Activity name
Activity status
Run start
Duration
Integration runtime
Copy data1 Succeeded 9/7/2023, 2:44:33 PM 9s AutoResolveIntegratior
Copy data1 Succeeded 9/7/2023, 2:44:33 PM 8s AutoResolveIntegratior
Copy data1 Succeeded 9/7/2023, 2:44:33 PM 9s AutoResolveIntegratior
ForEach1 Succeeded 9/7/2023, 2:44:33 PM 12s
Get Metadata1 Succeeded 9/7/2023, 2:44:29 PM 3s AutoResolveIntegratior


Microsoft Azure
Search resources, services, and docs (G+/)
Shellunext_1693422208...
UNEXT (NPUNEXT.ONMICROSOF...
Home > idashellsa03 | Containers >
foreachdata
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: foreachdata
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
OrderDetails.csv 9/7/2023, 2:44:39 PM Hot (Inferred) Block blob
Orders.csv 9/7/2023, 2:44:40 PM Hot (Inferred) Block blob
SalesTarget.csv 9/7/2023, 2:44:40 PM Hot (Inferred) Block blob

API Calls:





CDC (Change Data Capture):

We can use timestamp to note the times whenever the new records gets inserted/deleted/updated in the table. So now for only changed data we can perform the pipeline activity.