

## Custom Learnings

Day 8

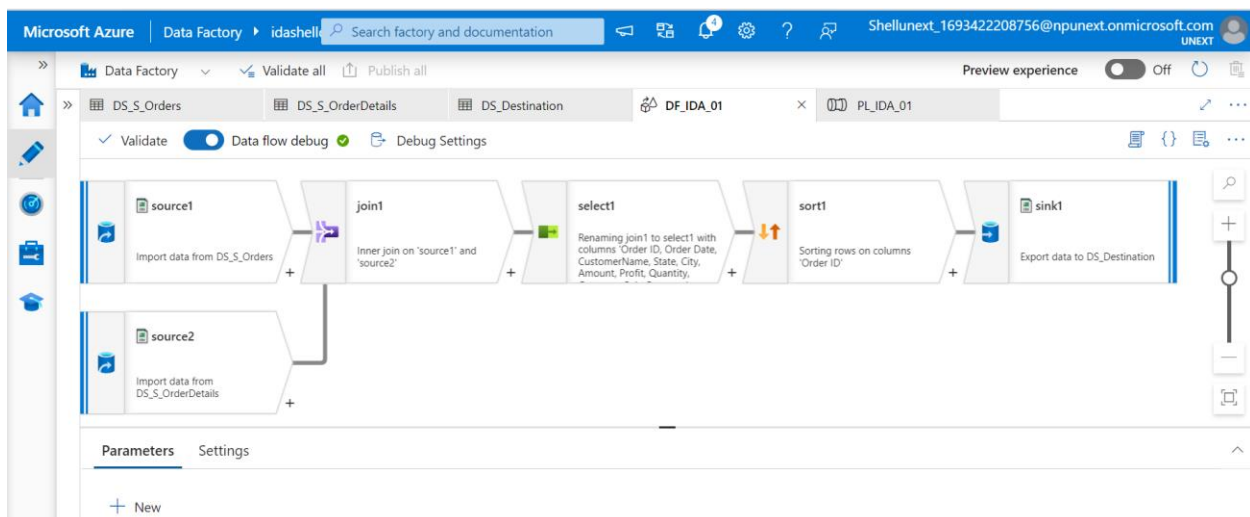
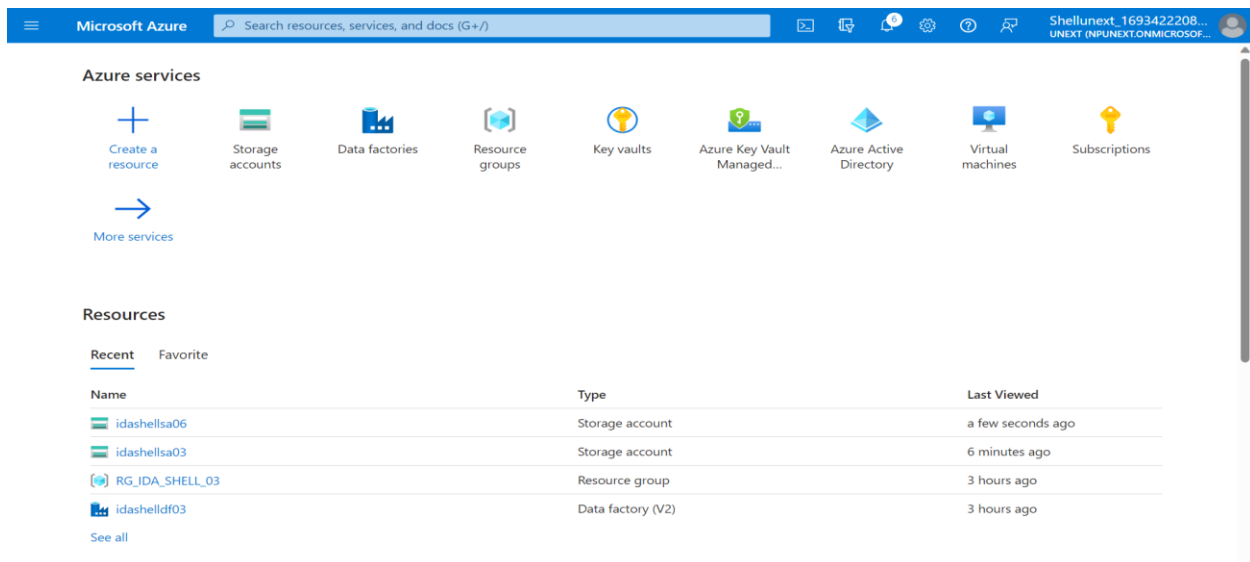
### Azure Data Factory

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:



Microsoft Azure | Data Factory | idashellsa03 | Search factory and documentation

Shellunext\_1693422208756@npunext.onmicrosoft.com UNEXT

Data Factory | Validate all | Publish all | Preview experience | Off

DS\_S\_Orders | DS\_S\_OrderDetails | DS\_Destination | DF\_IDA\_01 | PL\_IDA\_01

Validate | Debug | Add trigger | Data flow debug

Data flow

Data flow1

Parameters | Variables | Settings | Output

Pipeline run ID: 5de89e05-9484-4a8d-8e35-1db1aefc696c | Pipeline status: Succeeded | View debug run consumption

All status

Showing 1 - 1 of 1 items

Activity name	Activity status	Run start	Duration	Integration runtime	User properties	Activity run ID
Data flow1	Succeeded	9/7/2023, 12:00:29 PM	2m 24s	debugpool-8Cores-Gei		60be8120-7f44-426f-a090-fc

Microsoft Azure | Search resources, services, and docs (G+)

Shellunext\_1693422208... UNEXT (NPUNEXT.ONMICROSOFT...)

Home > idashellsa03 | Containers >

output Container

Search

Upload | Add Directory | Refresh | Rename | Delete | Change tier | Acquire lease | Break lease | Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: output

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type
transformed_output	9/7/2023, 12:01:49 PM	Hot (Inferred)		Block blob

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 | idashellsa03 | Search factory and documentation

Shellunext\_1693422208756@npunext.onmicrosoft.com UNEXT

Factory Resources

- Pipelines
  - 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

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	Us
Copy data1	Succeeded	9/7/2023, 2:44:33 PM	9s	AutoResolveIntegration	
Copy data1	Succeeded	9/7/2023, 2:44:33 PM	8s	AutoResolveIntegration	
Copy data1	Succeeded	9/7/2023, 2:44:33 PM	9s	AutoResolveIntegration	
ForEach1	Succeeded	9/7/2023, 2:44:33 PM	12s		
Get Metadata1	Succeeded	9/7/2023, 2:44:29 PM	3s	AutoResolveIntegration	

Microsoft Azure | Search resources, services, and docs (G+)

Shellunext\_1693422208... UNEXT (NPUNEXT.ONMICROSOFT...)

Home > idashellsa03 | Containers >

foreachdata

Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

- Shared access tokens
- Manage ACL
- Access policy
- Properties
- Metadata

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
<input type="checkbox"/> OrderDetails.csv	9/7/2023, 2:44:39 PM	Hot (Inferred)		Block blob
<input type="checkbox"/> Orders.csv	9/7/2023, 2:44:40 PM	Hot (Inferred)		Block blob
<input type="checkbox"/> SalesTarget.csv	9/7/2023, 2:44:40 PM	Hot (Inferred)		Block blob

API Calls:

Microsoft Azure | Data Factory | idashellsa03 | Search factory and documentation | Shellunext\_1693422208756@npunext.onmicrosoft.com | UNEXT

Validate all | Publish all | Preview experience | Off

**Factory Resources**

Filter resources by name

- Pipelines (3)
  - PL\_IDA\_03
  - PL\_IDA\_01
  - PL\_IDA\_02
- Change Data Capture (preview) (0)
- Datasets (7)
  - DS\_D\_FILENAME
  - DS\_Destination
  - DS\_JSON
  - DS\_S\_FILENAME
  - DS\_S\_OrderDetails
  - DS\_S\_Orders
  - DS\_Source\_for
- Data flows (1)
  - DF\_IDA\_01
- Power Query (0)

**PL\_IDA\_01** | **PL\_IDA\_02** | **DS\_Source\_for** | **DS\_S\_FILENAME** | **DS\_D\_FILENAME** | **DS\_Destination** | **DS\_JSON** | **DS\_S\_OrderDetails** | **DS\_S\_Orders** | **DS\_Source\_for**

Validate | Debug | Add trigger

Web (Web1) → Copy data (Copy data1)

Parameters | Variables | Settings | **Output**

**Pipeline run ID:** 5930c97d-f717-4d16-96c3-617e9bd10505 | **Pipeline status:** Succeeded | **Integration runtime:** AutoResolveIntegration

All status | Monitor in Azure Metrics | Export to CSV

Showing 1 - 2 of 2 items

Activity name	Activity status	Run start	Duration	Integration runtime	User
Copy data1	Succeeded	9/7/2023, 3:23:50 PM	10s	AutoResolveIntegration	
Web1	Succeeded	9/7/2023, 3:23:46 PM	3s	AutoResolveIntegration	

Microsoft Azure | Search resources, services, and docs (G+)

Home > idashellsa03 | Containers >

**output** | Container

Search | 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:** output

Search blobs by prefix (case-sensitive) | Show deleted objects

Name	Modified	Access tier	Archive status	Blob type
todos.txt	9/7/2023, 3:23:59 PM	Hot (Inferred)		Block blob
transformed_output	9/7/2023, 12:01:49 PM	Hot (Inferred)		Block blob

## 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.