Hands-On Lab: Union Transformation in Mapping Data Flow

Author: Sandeep Kumar Sharma

Learning Objective

This lab teaches you how to use the **Union Transformation** in Mapping Data Flow in Azure Data Factory. You will learn how to combine data from multiple datasets into a single output stream.

Learning Outcome

After completing this lab, you will be able to: - Understand the purpose of Union transformation - Create multiple source datasets - Configure a Union transformation - Preview and validate combined data

---

1. Introduction

The **Union Transformation** is used when you want to merge two or more datasets into a single output. It works similarly to the SQL **UNION ALL** operation, meaning it appends rows from multiple sources without removing duplicates.

In this lab, you will: - Create two CSV datasets - Load them into a Mapping Data Flow - Apply a Union transformation - Load the final combined data into Azure SQL Database

---

2. Datasets Used in This Lab

Dataset 1: `Employee_US.csv`

EmpID	EmpName	Country
1	John	USA
2	Robert	USA
3	Michael	USA

Dataset 2: `Employee_UK.csv`

EmpID	EmpName	Country
10	David	UK
11	George	UK
12	Mathew	UK

These two datasets will be merged using Union transformation.

---

3. Step-by-Step Hands-On Lab

Step 1: Upload CSV Files to Storage

1. Go to your Storage Account.
2. Create a folder named `employee/`.
3. Upload the two CSV files:
4. `Employee_US.csv`
5. `Employee_UK.csv`

Step 2: Create Two CSV Datasets in ADF

Dataset 1

1. Go to **Author** → **Datasets**.
2. Click **+ New Dataset** → CSV → ADLS/Blob.
3. Name it `EmployeeUS`.
4. Select the file path: `employee/Employee_US.csv`.
5. Save it.

Dataset 2

1. Create another dataset.
2. Name it `EmployeeUK`.
3. Select file: `employee/Employee_UK.csv`.
4. Save it.

Step 3: Create a Mapping Data Flow

1. Click **+** → **Data Flow**.

2. Choose **Mapping Data Flow**.
3. Name it: `UnionEmployeeDataFlow`.

Step 4: Add Source 1

1. Click **Add Source**.
2. Select dataset: `EmployeeUS`.
3. Name the source: `SourceUS`.
4. Enable **Allow schema drift**.

Step 5: Add Source 2

1. Click **+** on the canvas → **Add Source**.
2. Select dataset: `EmployeeUK`.
3. Name it: `SourceUK`.
4. Enable **Allow schema drift**.

Both sources now supply the employee data.

Step 6: Add Union Transformation

1. Click **+** next to any source.
2. Select **Union**.
3. In the Union settings:
4. Add both inputs: `SourceUS` and `SourceUK`.
5. Click **Add stream** if needed.
6. Ensure all column names match:
7. EmpID
8. EmpName
9. Country

If column names differ, apply **Select** transformation before Union.

Step 7: Add Sink Transformation (Azure SQL Database)

1. Click **+** after Union → **Sink**.
2. Choose SQL dataset: `EmployeeSQLTable`.
3. Create this SQL table beforehand:

```
CREATE TABLE EmployeesCombined (
    EmpID INT,
```

```
    EmpName VARCHAR(200),
    Country VARCHAR(50)
);
```

4. Enable **Auto Mapping**.

Step 8: Preview the Combined Data

1. Turn on **Data Flow Debug**.
2. Open **Data Preview** at Union.
3. You should now see:

EmpID	EmpName	Country
1	John	USA
2	Robert	USA
3	Michael	USA
10	David	UK
11	George	UK
12	Mathew	UK

Step 9: Run the Data Flow via Pipeline

1. Create a new pipeline.
2. Drag **Data Flow** activity.
3. Select `UnionEmployeeDataFlow`.
4. Validate → Debug.

Once executed, the SQL table will contain merged records from both CSV files.

---

4. Final Description

In this lab, you learned how the **Union Transformation** helps combine datasets in Azure Data Factory Mapping Data Flow. You successfully created two datasets, merged them, and loaded the combined output into SQL. This transformation is useful for combining data from regions, departments, or time-based batches.

This completes the hands-on exercise.