**Slowly Changing Dimension (SCD) Type 1 - Azure Data Factory Mapping Data Flow**

**Overview**

This document explains the structure and execution flow of the **SCD Type 1 Mapping Data Flow** used in Azure Data Factory (ADF). SCD Type 1 is used when historical data is not required, meaning that existing records are simply updated with new values.

SCD Type 1(Insert+Update)

**Why SCD Type 1?**

* + No need to maintain history.
  + Only the latest employee data is required.
  + **Overwrites existing records** instead of tracking changes.

**1. Data Sources**

* **Source 1 (source1)**
  + Reads employee data from a CSV file (PaperTowel2.csv) stored in an **Azure Data Lake Storage Gen2 (ADLS)** container (scdtype1 filesystem).
  + Schema includes:
    - ID (integer)
    - NAME (string)
    - CITY (string)
    - PHNO (long)

#### A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated

#### **Rename Columns (RenameColumns)**

* This transformation renames the source dataset fields by prefixing them with SRC\_ to avoid conflicts with the target table fields.

#### A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated**Derived Column (derivedColumn1)**

* Creates a new column SRC\_HASHKEY that generates a **hash key** using the CRC32 function.
* This hash key is used to detect changes in records.
* Formula
* SRC\_HASHKEY = crc32(concat(toString(SRC\_ID), SRC\_NAME, SRC\_CITY, toString(SRC\_PHNO)))

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated  **Target (Target)**

* Reads existing records from **Azure SQL Database** (EMPLOYEE\_SCDTYPE1 table) to check for updates.
* Schema includes:
  + EMP\_ID (integer)
  + EMP\_HASHKEY (long) – used for change detection.

#### A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated **Lookup (lookup1)**

* Joins the incoming source data (derivedColumn1) with the existing target table (Target) based on EMP\_ID.
* This helps determine whether a record already exists in the target table.

#### A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated

#### **Split (split1)**

* Divides records into two categories:
  + **Insert (New Records):** If EMP\_ID is NULL (not found in lookup), the record is new.
  + **Update (Modified Records):** If EMP\_ID exists but SRC\_HASHKEY is different from EMP\_HASHKEY, the record has changed and needs to be updated.
  + **Condition:**
  + **isNull(EMP\_ID), --- Insert**
  + **SRC\_ID == EMP\_ID && SRC\_HASHKEY != EMP\_HASHKEY--- Update**

#### A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated **Derived Column (derivedColumn2)** – For New Records

* Adds metadata for new records:

SRC\_CREATED\_BY = "dataflow"

SRC\_CREATED\_DATE = currentTimestamp()

SRC\_UPDATED\_BY = "dataflow"

#### SRC\_UPDATED\_DATE = currentTimestamp()A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated**Sink 1 (sink1) – Insert New Records**

* Inserts new records into EMPLOYEE\_SCDTYPE1 table.
* Schema Mapping:
* EMP\_ID = SRC\_ID
* EMP\_NAME = SRC\_NAME
* EMP\_CITY = SRC\_CITY
* EMP\_PHNO = SRC\_PHNO
* CREATED\_BY = SRC\_CREATED\_BY
* CREATED\_DATE = SRC\_CREATED\_DATE
* UPDATED\_BY = SRC\_UPDATED\_BY
* UPDATED\_DATE = SRC\_UPDATED\_DATE
* EMP\_HASHKEY = SRC\_HASHKEY
* **Insertable: True** (Only inserts, no updates or deletions)

#### A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated **Derived Column (derivedColumn3)** – For Updated Records

* Updates metadata for existing records:
* SRC\_UPDATED\_DATE = currentTimestamp()
* SRC\_UPDATED\_BY = "Dataflow Updated"

#### A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated **Alter Row (alterRow1)**

* Marks rows for updates based on a condition.
* **Condition:** Always updates existing records.

#### A screenshot of a computer Description automatically generated **Alter Row (alterRow1)**

* Marks rows for updates based on a condition.
* **Condition:** Always updates existing records. updateIf(1 == 1)

#### A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated **Sink 2 (sink2) – Update Existing Records**

* Updates existing records if changes are detected.
* Schema Mapping:
* EMP\_ID = SRC\_ID
* EMP\_NAME = SRC\_NAME
* EMP\_CITY = SRC\_CITY
* EMP\_PHNO = SRC\_PHNO
* UPDATED\_BY = SRC\_UPDATED\_BY
* UPDATED\_DATE = SRC\_UPDATED\_DATE
* EMP\_HASHKEY = SRC\_HASHKEY
* **Updateable: True** (Only updates, no inserts or deletions)
* **Primary Key:** EMP\_ID

### A screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a computer Description automatically generated **4. Execution Flow Summary**

1. **Extract data** from CSV (source1) and SQL table (Target).
2. **Rename columns** in the source data (RenameColumns).
3. **Generate hash keys** to track changes (derivedColumn1).
4. **Match records** with existing target data using a **Lookup** (lookup1).
5. **Split records** into:
   * **New Records (Insert Path)** → derivedColumn2 → sink1
   * **Updated Records (Update Path)** → derivedColumn3 → alterRow1 → sink2
6. **Load new data** into the target table (sink1).
7. **Update modified records** in the target table (sink2).

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated