DATA FLOW CONCEPT (USING SCD TYPE 1)

It will overwrite the existing data whatever changes happens(insert and update)

THE TRANSFORMATION PROPOSE

SELECT → IT WILL UPDATE THE COLUMN NAME, RENAME THE COLUMN, REMOVE UNNECESSARY COLUMNS

DERIVED COLUMN → IT WILL HELP TO GENERATE THE HASHKEY

LOOKUP →IT WILL HELP TO MERGE DATA LIKE LEFT JOIN

CONDITIONAL SPLIT → USED TO ROUTE DATA INTO DIFFERENT STREAMS BASED ON CONDITIONS

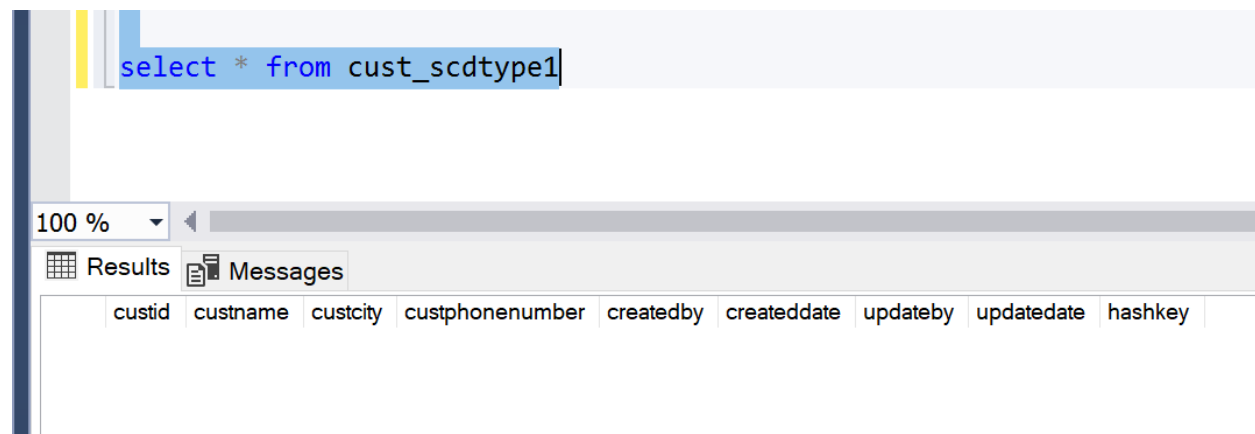ALTER ROW → USED TO DYNAMICALLY CONTROL ROW LEVEL ACTIONS (INSERT, UPDATE, DELETE, UPSERT(UPDATE+INSERT)

Create customer table

```sql
create table cust_scdtype1
(
custid int, custname varchar(100), custcity varchar(100), custphonenumber bigint,
createdby varchar(100), createddate datetime, updateby varchar(100), updatedate
datetime, hashkey bigint
)
```

```sql
select * from cust_scdtype1
```

100 %

Results | Messages

custid	custname	custcity	custphonenumber	createdby	createddate	updateby	updatedate	hashkey

Create a csv file in notepad

My day 1 file is

dataflowday1 - Notepad

File Edit Format View Help

```
id,name,city,phonenumber
1,abhi,hyd,1234567890
2,anna,cochine,0987654321
3,smith,blr,7856904523
4,jothi,tamilnadu,2435647898
```

My day 2 file is

dataflowday2 - Notepad

File Edit Format View Help

```
id,name,city,phonenumber
1,abhi,toronto,1234567890
5,suba,alberta,2345678901
```
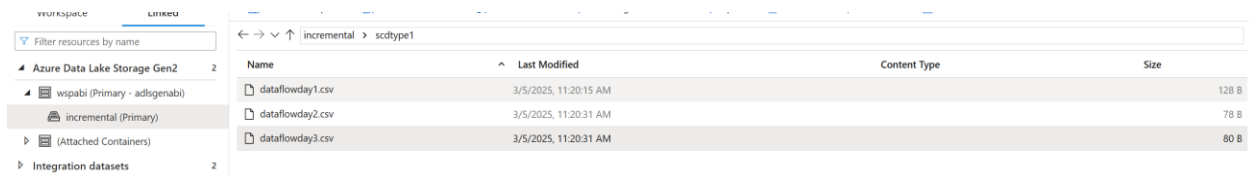
My day 3 file is

dataflowday3 - Notepad

File Edit Format View Help

```
id,name,city,phonenumber
1,abhi,toronto,1234567890
5,suba,tamilnadu,2354627666
```

Im gonna fetch the data from the csv to sqldb for that im gonna upload the file in my synapse workspace
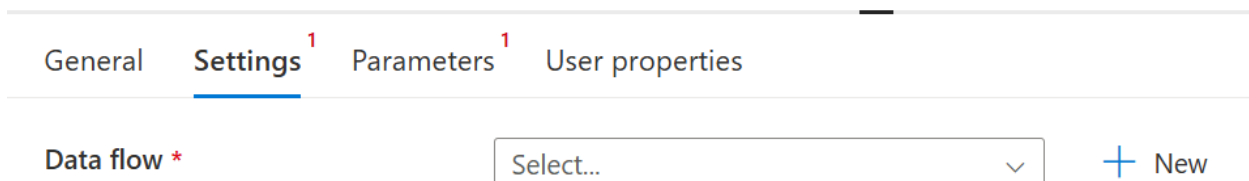

STEPS TO CREATE DATA FLOW IN SYNAPSE

STEP 1 : To create dataflow before starting the process click on data flow debug

STEP 2 : drag and drop the data flow and in setting →dataflow →click for new →give name for the data flow(sourcegen)

STEP 3 : My source is gen2 storage account which I have upload the 3 csv files

In sources setting →source type click for inline →inline data type change into delimited text → and in linked services give synapse storage account where you upload 3 csv files.

Source settings	Source options	Projection	Optimize	Inspect	Data preview ●

Output stream name *   sourcegen   Learn more ☐

Description   Add source dataset   ↻ Reset

Source type *

⊞ Integration dataset	◩ Inline	⬓ Workspace DB

Inline dataset type *   📄 DelimitedText ⌄

Linked service *   🖼 wspabi-WorkspaceDefaultStorage ⌄   ⚡ Test connection

Skip line count

Sampling * ⓘ   ◯ Enable   ⦿ Disable

STEP 4 : in source options →I gave my file path where my files are located

Source settings	Source options	Projection	Optimize	Inspect	Data preview ●

∨ **File settings**

File mode ⓘ   ⦿ File   ◯ Wildcard

File path *   incremental / scdtype1 / dataflowday1.csv   📁 Browse

STEP 5: In source options check first row as header

First row as header   ✅

STEP 6 : In projection → click on import the schema

Source settings	Source options	Projection	Optimize	Inspect	Data preview ●

←┘ Import schema   ✕ Clear schema   📇 Schema options

Column name	Type	Format
id	12ₛ short	Specify format ⌄
name	abc string	Specify format ⌄
city	abc string	Specify format ⌄
phonenumber	12ₗ long	Specify format ⌄

STEP 7 : Click on data preview (I have imported day1 csv file)

	id	name	city	phonenumber	
+	1	abhi	hyd	1234567890	
+	2	anna	cochine	987654321	
+	3	smith	blr	7856904523	
+	4	jothi	tamilnadu	2435647898	

Source settings   Source options   Projection   Optimize   Inspect   **Data preview** ●

Number of rows + INSERT 4   UPDATE 0   × DELETE 0   UPSERT 0   LOOKUP 0   ❌ ERROR 0   TOTAL 4

Refresh | ∨   Typecast ∨   Modify ∨   Map drifted   Statistics × Remove   ↓ Export to CSV | ∨

STEP 8 : By clicking + in the sourcesgen Im gonna start new transformation called select

In select transformation I wanna rename the column name by doing this step

Step1 : delete all the column in here

	sourcegen's column		Name as			
	12s id	∨	→	id	+	🗑
	abc name	∨	→	name	+	🗑
	abc city	∨	→	city	+	🗑
	12l phonenumber	∨	→	phonenumber	+	🗑

Step 2 : to do this delete all the column and click on + symbol click on rule-bases mapping.

==Concat('src_',$$)==

By doing this it will change all the column name into src_id,src_name,src_city,src_phonenumber

	sourcegen's column			Name as				
	1==1	×ᵥ	≫	→	concat('src_',$$)	abc	+	🗑 ∞

STEP 9 : click on preview data and refresh (column name changed)

Number of rows + INSERT 4   UPDATE 0   × DELETE 0   UPSERT 0   LOOKUP 0   ❌ ERROR 0   TOTAL 4

Refresh | ∨   Typecast ∨   Modify ∨   Map drifted   Statistics × Remove   ↓ Export to CSV | ∨

	src_id	src_name	src_city	src_phonenumber	
+	1	abhi	hyd	1234567890	
+	2	anna	cochine	987654321	
+	3	smith	blr	7856904523	
+	4	jothi	tamilnadu	2435647898	

STEP 10 : By clicking + in the **select**(rename) Im gonna start new transformation called drivedcolumn
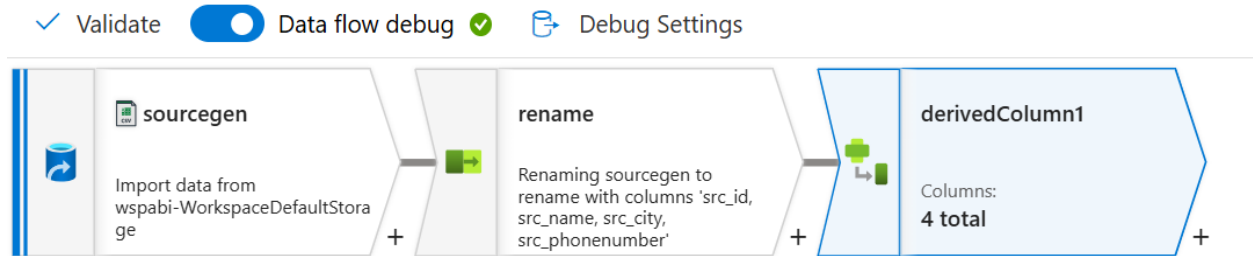
STEP 11 : in derived column , have to create a new column (hashkey) and use concat expressions

crc32(concat(toString(src_id),src_name,src_city,toString(src_phonenumber)))

STEP 12 : I wanna add my target (sink)

In source setting →source type →inline →inline dataset type →sql db →linked services →servername (ssms)
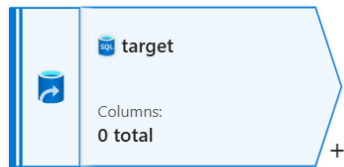
Source settings | Source options | Projection | Optimize | Inspect | Data preview ●

Output stream name *	target	Learn more ⬀
Description	Add source dataset	↻ Reset
Source type *	▦ Integration dataset ▨ Inline ◯ Workspace DB	
Inline dataset type *	▦ Azure SQL Database ⌄	
Linked service *	▦ azsqlserver ⌄	⚡ Test connection ✎ Edit ＋ New
Sampling * ⓘ	◯ Enable ◉ Disable	

STEP 13 : I wanna compare this 2 column whether this is existing or not

<mark>select custid, hashkey from cust_scdtype1</mark>

target

Columns:
0 total

+

| Source settings | **Source options** | Projection | Optimize | Inspect | Data preview ● |

Input   ○ Table ● Query ○ Stored procedure

Query * ⓘ   select custid, hashkey from cust_scdtype1   ✏️

Incremental column ⓘ   ☐

Isolation level ⓘ   Read uncommitted   ⌄

STEP 14 : click on projection →click on import schema

| Source settings | Source options | **Projection** | Optimize | Inspect | Data preview ● |

←| Import schema   ✕ Clear schema   ▤ Schema options   ✏️ Overwrite schema

Column name	↑↓ ▽ Type
custid	123 integer
hashkey	12l long

STEP 15: By clicking + in the derived(hashkey)→ start new transformation called lookup
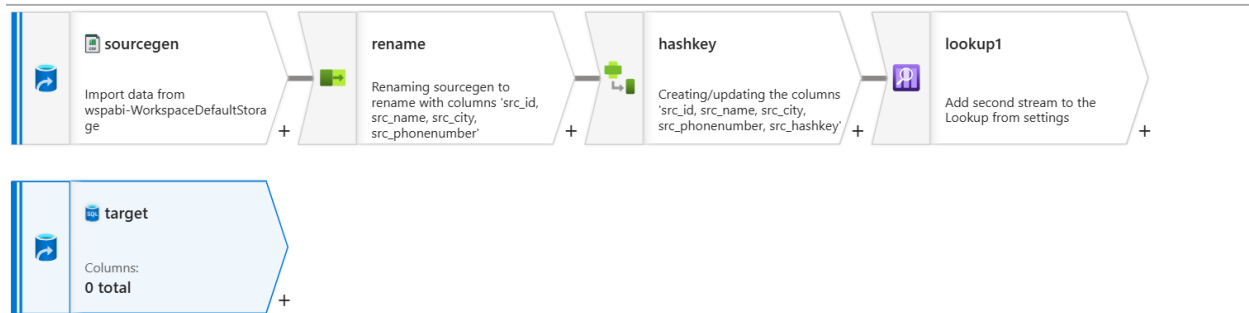
STEP 16 : in lookup setting →lookup stream (select target (sql db) and in lookup conditions → select src_id for left and custid for right

STEP 17 : in data preview→refresh

In the above picture I can only see the src_id,src_name,src_city,src_phonenumber this is my csv file and in the custid,hashkey is null(I don't inserted data in the sqldb), to bring data use conditional split

STEP 18 : By clicking + in the lookup→ start new transformation called conditional split
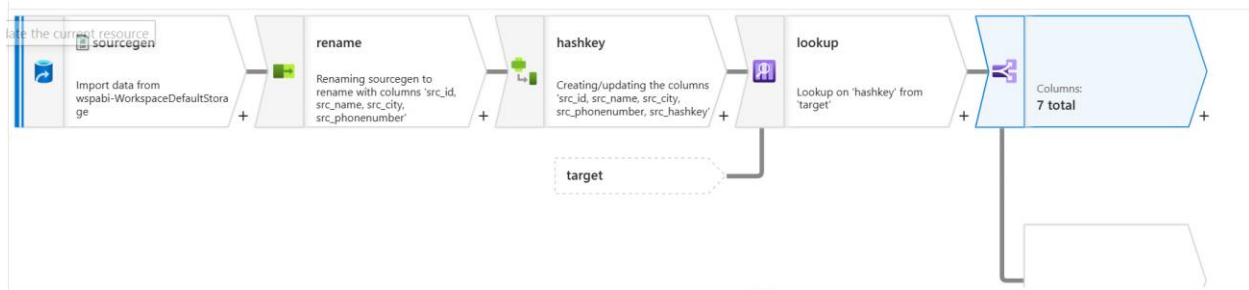
STEP 19 : delete the default stream and add the new (add conditional stream)

STEP 20 : if the custid is null consider this as a new record

isNull(custid)

If the custid is not null consider this as a existing record and update this

Hashkey is already generated and it will generate another hashkey use the below code
code

src_id==custid && src_hashkey!=hashkey

Output stream name *	split	Learn more ⬚
Description	Conditionally distributing the data in custid, src_id, custid, src_hashkey, hashkey groups, based on columns '{1}'	↻ Reset
Incoming stream *	lookup	
Split on	◉ First matching condition ○ All matching conditions	

Split condition

Stream names	Condition	
insert	isNull(custid)	⌄ + ⬚
update	src_id==custid && src_hashkey!=hashkey	⌄ + ⬚

STEP 21 : By clicking + in the insert Im gonna start new transformation called derived column (to insert audit column)
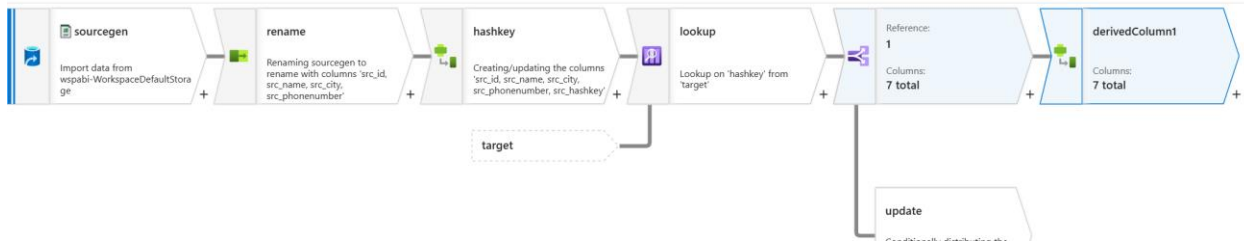
STEP 22 : I wanna insert audit column (createdby, createddate, updateby, updatedate)

Add →new columns

Output stream name *	insertauditcolumn	Learn more ⬚
Description	Creating/updating the columns 'src_id, src_name, src_city, src_phonenumber, src_hashkey, custid, hashkey,	↻ Reset
Incoming stream *	split@insert	

+ Add   ⧉ Clone   🗑 Delete   ⬚ Open expression builder

Columns * ⓘ

	Column	Expression	
☐	src_createdby ⌄	Enter expression... ANY	+ 🗑
☐	src_createddate ⌄	Enter expression... ANY	+ 🗑
☐	src_updateby ⌄	Enter expression... ANY	+ 🗑
☐	src_updatedate ⌄	Enter expression... ANY	+ 🗑
	No results found		

Add →expression aswell

Column	Expression
src_createdby	'dataflow'
src_createddate	currentTimestamp()
src_updateby	'dataflow'
src_updatedate	currentTimestamp()

Derived column's settings   Optimize   Inspect   Data preview ●

Output stream name *   insertauditcolumn   Learn more ↗

Description   Creating/updating the columns 'src_id, src_name, src_city, src_phonenumber, src_hashkey, custid, hashkey,   ↻ Reset

Incoming stream *   split@insert

Columns * ⓘ   + Add   🗐 Clone   🗑 Delete   ⤢ Open expression builder

STEP 23 : By clicking + in the derived (audit) add new transformation called sink
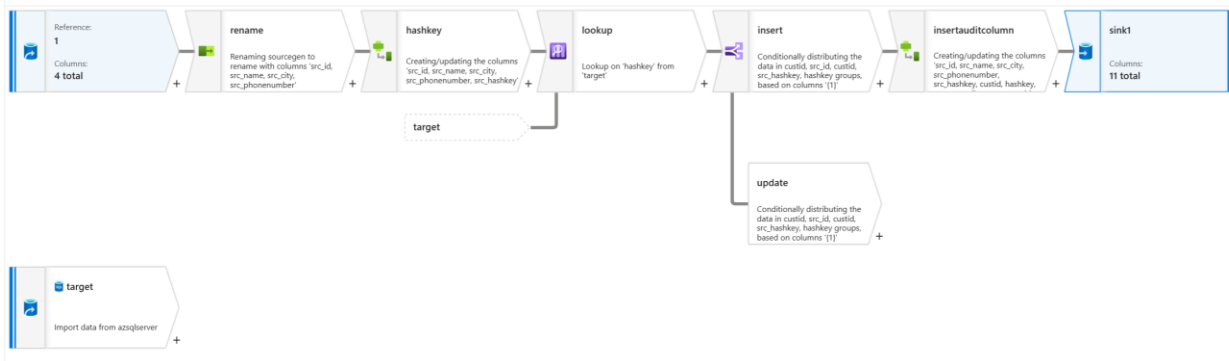
STEP 23 : In sink →sink type click inline → inline dataset type click sqldb →in linked services click your ssms(server)

Sink   Settings   Errors   Mapping   Optimize   Inspect   Data preview ●

Output stream name *   [sink]   Learn more ⬀

Description   [Add sink dataset   ]   ↻ Reset

Incoming stream *   [insertauditcolumn   ⌄]

Sink type *   ┌──────────┬──────────┬──────────┬──────────┐
                            │ ⊞   │ ▨   │ 🗄   │ 🗐   │
                            │Integration│ Inline │Workspace DB│ Cache │
                            │ dataset │   │   │   │
                            └──────────┴──────────┴──────────┴──────────┘

Inline dataset type *   [📘 Azure SQL Database   ⌄]

Linked service *   [📘 azsqlserver   ⌄]   ✎ Test connection   ✎ Edit   ＋ New

Options   ☑ Allow schema drift ⓘ

                            ☐ Validate schema ⓘ

STEP 24 : In setting if you unable to see the scheman refresh

Sink   **Settings**   Errors   Mapping   Optimize   Inspect   Data preview ●

Schema name *   [Click refresh to load schema options   ⌄]   ↻ Refresh   ✅ Success

STEP 25 : update the schema and table name and uncheck the tempdb

Sink   **Settings**   Errors   Mapping   Optimize   Inspect   Data preview ●

Schema name *   [dbo   ⌄]   ↻ Refresh   ✅ Success

Table name *   [cust_scdtype1   ⌄]

Table action   ⦿ None   ◯ Recreate table   ◯ Truncate table

Update method ⓘ   ☑ Allow insert

                            ☐ Allow delete

                            ☐ Allow upsert

                            ☐ Allow update

Use tempdb ⓘ   ☐

Interim table schema   [Click refresh to load schema options   ⌄]   ↻ Refresh

STEP 26 : in mapping →uncheck the auto mapping and click →importschema

Once the importschema inserted click on reset

| Sink | Settings | Errors | **Mapping** | Optimize | Inspect | Data preview ● |

Options ☑ Skip duplicate input columns ⓘ
☑ Skip duplicate output columns ⓘ

☐ Auto mapping ⓘ ＋ Add mapping 🗑 Delete ↻ Reset ↩ Import schema 🔧 View schema

☐	Input columns	🔽			Output columns	🔽		
☐	123 custid	⌄	→		123 custid	⌄	🗑	＋
☐		⌄	→		abc custname	⌄	🗑	＋
☐		⌄	→		abc custcity	⌄	🗑	＋
☐		⌄	→		121 custphonenumber	⌄	🗑	＋
☐		⌄	→		abc createdby	⌄	🗑	＋
☐		⌄	→		🕐 createddate	⌄	🗑	＋
☐		⌄	→		abc updateby	⌄	🗑	＋
☐		⌄	→		🕐 updatedate	⌄	🗑	＋
☐	121 hashkey	⌄	→		121 hashkey	⌄	🗑	＋

STEP 27 : give the values

| Sink | Settings | Errors | **Mapping** | Optimize | Inspect | Data preview ● |

⚠ At least one incoming column is mapped to a column in the sink dataset schema with a conflicting type, which can cause NULL values or runtime errors.

Options ☑ Skip duplicate input columns ⓘ
☑ Skip duplicate output columns ⓘ

☐ Auto mapping ⓘ ＋ Add mapping 🗑 Delete ↻ Reset ↩ Import schema 🔧 View schema

☐	Input columns	🔽			Output columns	🔽		
☐	12s src_id	⌄	→		123 custid	⌄	🗑	＋
☐	abc src_name	⌄	→		abc custname	⌄	🗑	＋
☐	abc src_city	⌄	→		abc custcity	⌄	🗑	＋
☐	121 src_phonenumber	⌄	→		121 custphonenumber	⌄	🗑	＋
☐	abc src_createdby	⌄	→		abc createdby	⌄	🗑	＋
☐	🕐 src_createddate	⌄	→		🕐 createddate	⌄	🗑	＋
☐	abc src_updateby	⌄	→		abc updateby	⌄	🗑	＋
☐	🕐 src_updatedate	⌄	→		🕐 updatedate	⌄	🗑	＋
☐	121 src_hashkey	⌄	→		121 hashkey	⌄	🗑	＋

STEP 28 : click on data preview

| Number of rows | + INSERT | 0 | • UPDATE | 0 | × DELETE | 0 | •+ UPSERT | 0 | LOOKUP | 0 | ✖ ERROR | 0 | TOTAL | 4 |

Refresh | ∨   Statistics   ⬇ Export to CSV | ∨

	custid		custname		custcity		custphonenumber		createdby		createddate		updateby		updatedate		hashkey	
•	1		abhi		hyd		1234567890		dataflow		2025-03-05 18:23:18.449		dataflow		2025-03-05 18:23:18.449		3602628959	
•	2		anna		cochine		987654321		dataflow		2025-03-05 18:23:18.449		dataflow		2025-03-05 18:23:18.449		57269329	
•	3		smith		blr		7856904523		dataflow		2025-03-05 18:23:18.449		dataflow		2025-03-05 18:23:18.449		2823754298	
•	4		jothi		tamilnadu		2435647898		dataflow		2025-03-05 18:23:18.449		dataflow		2025-03-05 18:23:18.449		1550425258	

STEP 30 : By clicking + in the update Im gonna start new transformation called derived column (to update audit column)
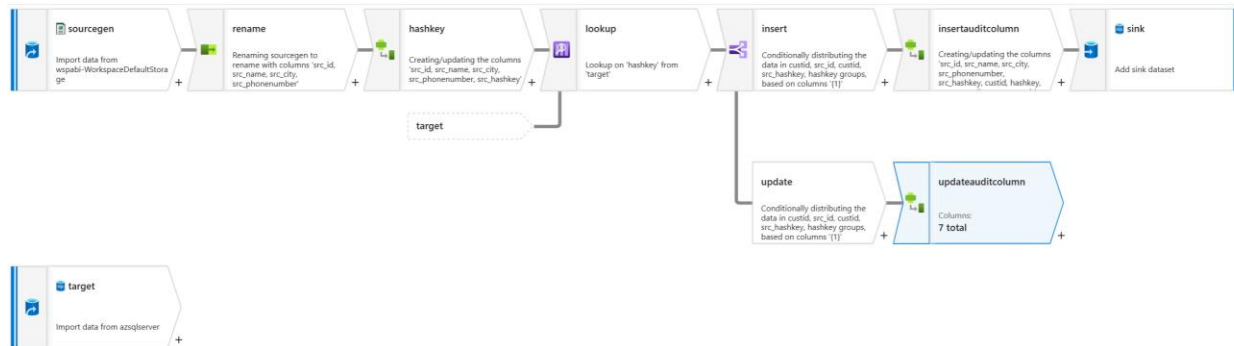
STEP 31 : in add column just give the updateby and update date

STEP 32 : By clicking + in the derivedcolumn Im gonna start new transformation called alter row
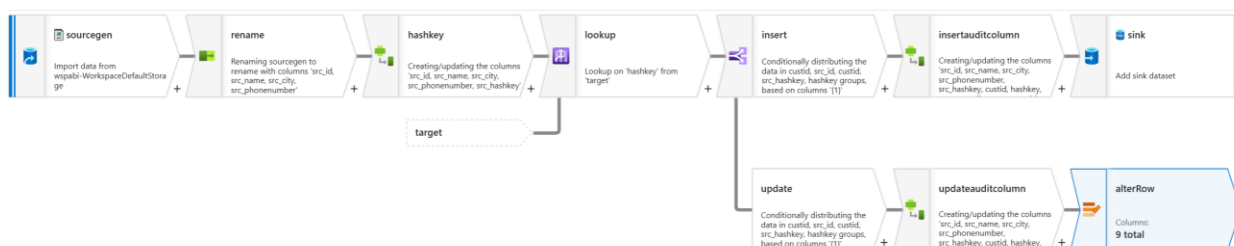
STEP 33 : Permission to modify the data

Alter row settings   Optimize   Inspect   Data preview ●

Output stream name *	alterRow	Learn more ↗
Description	Add expressions to alter rows	⟳ Reset
Incoming stream *	updateauditcolumn ⌄	

Alter row conditions * ⓘ

✳ Update if ⌄	1==1

Open expression builder

Updateif   1==1
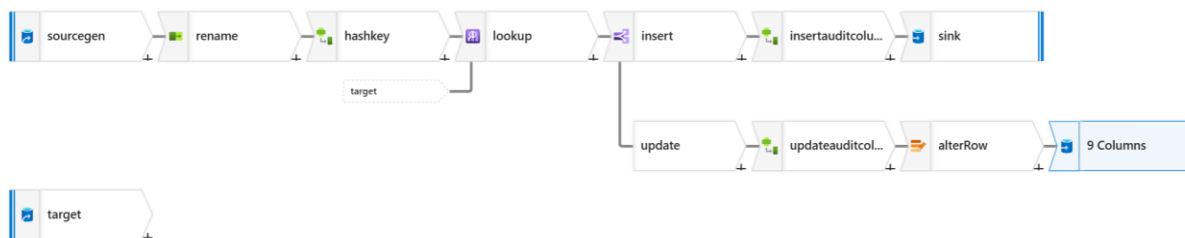
STEP 34 : By clicking + in the alterrow Im gonna start new transformation called sink

STEP 35 : give sql db and server name (ssms   )

| Sink | Settings | Errors | Mapping | Optimize | Inspect | Data preview ● |

| Output stream name * | sink1 | Learn more ⬈ |

| Description | Add sink dataset | ↺ Reset |

| Incoming stream * | alterRow ⌄ |

| Sink type * | ▦ Integration dataset | ◩ Inline | 🛢 Workspace DB | 🗄 Cache |

| Inline dataset type * | 🔲 Azure SQL Database ⌄ |

| Linked service * | 🔲 azsqlserver ⌄ | 🔌 Test connection | ✏ Edit | + New |

| Options | ☑ Allow schema drift ⓘ |
| | ☐ Validate schema ⓘ |

STEP 36 : in setting →give schema and table name

| Sink | **Settings** | Errors | Mapping | Optimize | Inspect | Data preview ● |

| Schema name * | dbo ⌄ | ↺ Refresh |

| Table name * | cust_scdtype1 ⌄ |

| Table action | ⦿ None | ○ Recreate table | ○ Truncate table |

Update method ⓘ	☐ Allow insert
	☐ Allow delete
	☐ Allow upsert
	☑ Allow update

| Skip writing key columns ⓘ | ☐ |

STEP 37 : in mapping →uncheck auto mapping →click on importschema and click reset

STEP 38 : delete the createby and createdate I only need updateby and updatedate

STEP 39 : in setting give key column as custid (is my keycolumn based on custid I will update data)

Sink   **Settings**   Errors   Mapping   Optimize   Inspect   Data preview   ●

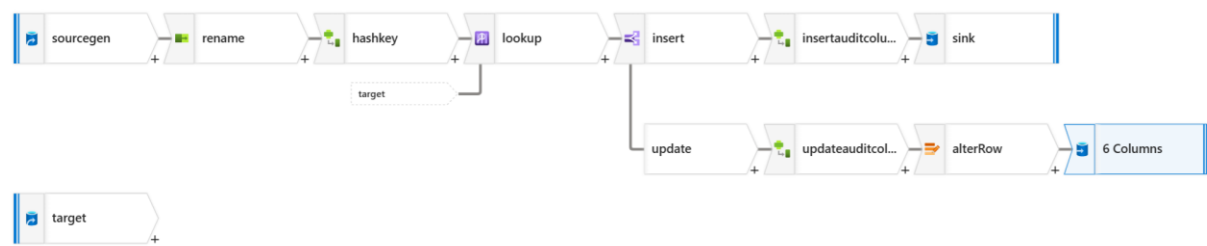☐ Allow upsert

✓ Allow update

Skip writing key columns ⓘ   ☐

Key columns * ⓘ   ⦿ List of columns   ○ Custom expression ⓘ

| custid | 123 |

＋   🗑

STEP 40 : published successfully (THIS IS THE FLOW)

STEP 41 : pipeline debug

Parameters   Variables   Settings   **Output**   ⌃

Pipeline run ID: 63da5865-099e-42c1-9277-506fb0c10274 [@] ↻ ⓘ   **Pipeline status** ✓ Succeeded   View debug run consumption

All status ⌄   Monitor in Azure Metrics ⬈   ⤓ Export to CSV | ⌄

Showing 1 - 1 of 1 items

Activity name ↑↓	Activity st... ↑↓	Activit... ↑↓	Run start ↑↓	Duration ↑↓	Integration runtime ↑↓
Data flow	✓ Succeeded	Data flow	3/5/2025, 3:38:17 PM	1m 22s	AutoResolveIntegrationRuntime (Australi

STEP 44 : output in ssms

```sql
select * from cust_scdtype1
```

100 %

Results | Messages

	custid	custname	custcity	custphonenumber	createdby	createddate	updateby	updatedate	hashkey
1	1	abhi	toronto	1234567890	dataflow	2025-03-05 20:38:36.983	dataflow-updated	2025-03-05 20:42:00.043	4224953451
2	3	smith	blr	7856904523	dataflow	2025-03-05 20:38:36.983	dataflow	2025-03-05 20:38:36.983	2823754298
3	4	jothi	tamilnadu	2435647898	dataflow	2025-03-05 20:38:36.983	dataflow	2025-03-05 20:38:36.983	1550425258
4	2	anna	cochine	987654321	dataflow	2025-03-05 20:38:36.983	dataflow	2025-03-05 20:38:36.983	57269329
5	5	suba	alberta	2345678901	dataflow	2025-03-05 20:42:08.380	dataflow	2025-03-05 20:42:08.380	1465066377