Hands-On Lab 7: Surrogate Keys & Advanced Derived Columns in Azure Data Factory

Lab Created by : Dr. Sandeep Kumar Sharma

Concept Overview (Before HOL)

Surrogate keys and advanced derived expressions are crucial in ETL pipelines.

Why Surrogate Keys?

- Natural keys (like customer_id) may not be reliable
- Surrogate keys ensure unique, consistent identifiers
- Used widely in **Data Warehousing (Dimension tables)**

Why Advanced Derived Columns?

Used for: - Data standardization (uppercase, clean strings) - Business logic (calculations, flags) - Timestamps & audit columns

This lab teaches BOTH concepts with a simple dataset.

What This Lab Will Do

You will: 1. Generate **surrogate keys** using `surrogateKey()` 2. Add **audit columns** (load_time) 3. Apply string functions like `upper()`, `trim()` 4. Apply conditional expressions (`iif`)

Step 1 — Upload File

Upload this file into `lab7/`.

customers_raw.csv

```
customer_name,city,age
Sandeep ,Delhi,34
  Varun,Pune,28
Meena,Mumbai,42
```

Step 2 — Create Dataset

Dataset name: `ds_customers_lab7` File: `lab7/customers_raw.csv` Format: CSV, header = true

Step 3 — Create Mapping Data Flow

Name: **df_surrogate_advanced_lab7**

Add Source

- Name: `src_customers`
- Dataset: `ds_customers_lab7`

Part A — Generate Surrogate Key

1. Add **Derived Column** transformation
2. Add new column:
3. Name: `customer_sk`
4. Expression: `surrogateKey()`

This generates unique incremental numbers:

```
1
2
3
```

Part B — Apply Advanced Transformations

In the same Derived Column transformation, add:

Clean customer_name

```
clean_name = trim(upper(customer_name))
```

Flag senior customers (> 40)

```
senior_flag = iif(age > 40, 'YES', 'NO')
```

Add load timestamp

```
load_time = currentTimestamp()
```

Step 4 — Add Sink

Dataset: `lab7/output/clean_customers/` Format: CSV

Step 5 — Debug & Run

1. Enable Debug
2. Inspect surrogate keys and transformations
3. Publish & Trigger

Expected Output

```
customer_sk,clean_name,city,age,senior_flag,load_time
1,SANDEEP,Delhi,34,NO,2025-01-01T12:00:00Z
2,VARUN,Pune,28,NO,2025-01-01T12:00:00Z
3,MEENA,Mumbai,42,YES,2025-01-01T12:00:00Z
```

Lab Completed

This lab covered: ✔️Surrogate Keys via surrogateKey()
✔️Advanced Derived Columns
✔️Audit columns & Data standardization