

ITMD 526 – Data warehousing

Assignment 3

Nikhitha Kamath – A20473286

Importing Excel File

Microsoft Excel input

Step name: excellInput_Ni

#	Name	Type	Length	Precision	Trim type
1	customer_id	String			none
2	first_name	String			none
3	last_name	String			none
4	date_of_birth	Date			none
5	city	String			none
6	state	String			none
7	effective_date	Date			none
8					

Rows of step: excellInput_Nikhitha (5 rows)

#	customer_id	first_name	last_name	date_of_birth	city	state	effective_date
1	C1050	BethAnn	Cox	1993/01/03	New York	New York	2015/01/01
2	C1197	Panpan	Bressler	1992/01/30	PHILADELPHIA	Pennsylvania	2015/01/01
3	C16400	Tairan	Adelson	1993/06/30	Pittsburgh	Pennsylvania	2015/01/01
4	C16474	Marco	Hussie	1971/01/29	Pittsburgh	Pennsylvania	2015/01/01
5	C1050	Ann	Benson	1993/01/03	Chicago	Illinois	2017/08/01

Sorting rows by **customer_id** and then by **effective_date**

Sort rows

Step name: Sort rows

Sort directory: %%java.io.tmpdir%%

TMP-file prefix: out

Sort size (rows in memory): 1000000

Free memory threshold (in %):

Compress TMP Files? ☒

Only pass unique rows? (verifies keys only) ☐

Fields:

#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?
1	customer_id	Y	N	N	0	N
2	effective_date	Y	N	N	0	N

Dimension Lookup

1. Connect to a database
2. Give a target table name
3. 'Punch Through' Values. This overwrites the most recent data. This is SCD Type 1

- first_name: punch through; this is a scd type 1 variable.
- last_name: punch through; this is a scd type 1 variable.
- date_of_birth: punch through; this is a scd type 1 variable.
- city: punch through; this is a scd type 1 variable.
- State: punch through; this is a scd type 1 variable.

The screenshot shows the 'Dimension lookup/update' configuration window. The 'Step name' is 'dimCustomer_Nikhitha'. The 'Update the dimension?' checkbox is checked. The 'Connection' is 'assign3_Nikhitha'. The 'Target schema' is 'assign3_Nikhitha'. The 'Target table' is 'target_nikhitha'. The 'Commit size' is '100'. The 'Enable the cache?' checkbox is checked. The 'Pre-load the cache?' checkbox is unchecked. The 'Cache size in rows (0 = cache all)' is '5000'.

The 'Keys' tab is selected, showing the 'Lookup/Update fields' table:

#	Dimension field	Stream field to compare with	Type of dimension update
1	first_name	first_name	Punch through
2	last_name	last_name	Punch through
3	date_of_birth	date_of_birth	Punch through
4	city	city	Punch through
5	state	state	Punch through

The 'Technical key field' is 'customer_skey'. The 'Creation of technical key' options are: 'Use table maximum + 1' (unchecked), 'Use sequence' (unchecked), and 'Use auto increment field' (checked). The 'Version field' is 'version'. The 'Stream Datefield' is 'effective_date'. The 'Date range start field' is 'date_from' with 'Min. year' '1900'. The 'Use an alternative start date?' checkbox is unchecked. The 'Table date range end' is 'date_to' with 'Max. year' '9998'.

Buttons: OK, Cancel, Get Fields, SQL, Help.

Successful Execution of Transformation and Job

The screenshot shows the 'Execution Results' window for the job 'dimCustomer_Nikhitha'. The job flow is: 'excellInput_Nikhitha' -> 'Sort rows' -> 'dimCustomer_Nikhitha'. All steps are marked with green checkmarks, indicating successful execution.

The 'Execution Results' tab is selected, showing the following log entries:

```
2021/10/21 11:51:12 - Spoon - Using legacy execution engine
2021/10/21 11:51:12 - Spoon - Transformation opened.
2021/10/21 11:51:12 - Spoon - Launching transformation [slowly-changing-dimension]...
2021/10/21 11:51:12 - Spoon - Started the transformation execution.
2021/10/21 11:51:12 - slowly-changing-dimension - Dispatching started for transformation [slowly-changing-dimension]
2021/10/21 11:51:12 - excellInput_Nikhitha.0 - Finished processing (I=5, O=0, R=0, W=5, U=0, E=0)
2021/10/21 11:51:12 - Sort rows.0 - Finished processing (I=0, O=0, R=5, W=5, U=0, E=0)
2021/10/21 11:51:12 - dimCustomer_Nikhitha.0 - Finished processing (I=5, O=4, R=5, W=5, U=2, E=0)
2021/10/21 11:51:12 - Spoon - The transformation has finished!!
```

Execution Results

Logging | History | Job metrics | Metrics

2021/10/21 11:51:31 - scd-main - Starting entry [scd]
2021/10/21 11:51:31 - scd - Using run configuration [Pentaho local]
2021/10/21 11:51:31 - scd - Using legacy execution engine
2021/10/21 11:51:31 - slowly-changing-dimension - Dispatching started for transformation [slowly-changing-dimension]
2021/10/21 11:51:32 - excellInput_Nikhitha.0 - Finished processing (I=5, O=0, R=0, W=5, U=0, E=0)
2021/10/21 11:51:32 - Sort rows.0 - Finished processing (I=0, O=0, R=5, W=5, U=0, E=0)
2021/10/21 11:51:32 - dimCustomer_Nikhitha.0 - Finished processing (I=4, O=0, R=5, W=5, U=4, E=0)
2021/10/21 11:51:32 - scd-main - Finished job entry [scd] (result=[true])
2021/10/21 11:51:32 - scd-main - Finished job entry [truncateTable_Nikhitha] (result=[true])
2021/10/21 11:51:32 - scd-main - Job execution finished
2021/10/21 11:51:32 - Spoon - Job has ended.

SCD Type 1 Output

All columns were punch through SCD type 1 variables, therefore the details were overwritten.

customer_key	version	date_from	date_to	customer_id	first_name	last_name	date_of_birth	city	state
1	1	(NULL)	(NULL)	(NULL)	OK	(NULL)	OK	(NULL)	OK
2	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C1050	5B Ann	3B Benson	6B 1993-01-03 00:00:00	Chicago 7B	Illi... 8B
3	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C1197	5B Panpan	6B Bressler	8B 1992-01-30 00:00:00	PHIL... 12B	Penn... 12B
4	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C16400	6B Tairan	6B Adelson	7B 1993-06-30 00:00:00	Pitt... 10B	Penn... 12B
5	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C16474	6B Marco	5B Hussie	6B 1971-01-29 00:00:00	Pitt... 10B	Penn... 12B

	A	B	C	D	E	F	G	H	I
1	customer_id	first_name	last_name	date_of_birth	city	state	effective_date		
2	C1050	BethAnn	Cox	1/3/1993	New York	New York	1/1/2015		
3	C1050	Ann	Benson	1/3/1993	Chicago	Illinois	8/1/2017		
4	C1197	Panpan	Bressler	1/30/1992	PHILADELPHIA	Pennsylvania	1/1/2015		
5	C16400	Tairan	Adelson	6/30/1993	Pittsburgh	Pennsylvania	1/1/2015		
6	C16474	Marco	Hussie	1/29/1971	Pittsburgh	Pennsylvania	1/1/2015		
7									
8									
9									
10									
11									
12									
13									
14									
15	customer_key	version	customer_id	first_name	last_name	date_of_birth	city	state	
16	1	1							
17	2	1	C1050	Ann	Benson	1993-01-03 00:00:00	Chicago	Illinois	
18	3	2	C1050	Ann	Benson	1993-01-03 00:00:00	Chicago	Illinois	
19	4	1	C1197	Panpan	Bressler	1992-01-30 00:00:00	PHILADELPHIA	Pennsylvania	
20	5	1	C16400	Tairan	Adelson	1993-06-30 00:00:00	Pittsburgh	Pennsylvania	
21	6	1	C16474	Marco	Hussie	1971-01-29 00:00:00	Pittsburgh	Pennsylvania	
22									
23									

SCD Type 2

1. Connect to a database
 2. Give a target table name
 3. 'Insert' Values. Adds a new row. This is SCD Type 2
- first_name: insert; this is a scd type 2 variable.
 - last_name: punch through; this is a scd type 1 variable.
 - date_of_birth: punch through; this is a scd type 1 variable.
 - city: punch through; this is a scd type 1 variable.
 - State: punch through; this is a scd type 1 variable.

Dimension lookup/update

Step name: dimCustomer_Nikhitha

Update the dimension? ☒

Connection: assign3_Nikhitha

Target schema: target_nikhitha

Target table: target_nikhitha

Commit size: 100

Enable the cache? ☒

Pre-load the cache? ☐

Cache size in rows (0 = cache all): 5000

Keys Fields

#	Dimension field	Stream field to compare with	Type of dimension update
1	first_name	first_name	Insert
2	last_name	last_name	Punch through
3	date_of_birth	date_of_birth	Punch through
4	city	city	Punch through
5	state	state	Punch through

Technical key field: customer_key

Version field: version

Stream Datefield: effective_date

Date range start field: date_from

Date range end field: date_to

Min. year: 1900

Max. year: 9998

Use an alternative start date? ☐ Select Option

Use auto increment field ☒

OK Cancel Get Fields SQL

SCD Type 2 Output

1 Result 2 Profiler 3 Messages 4 Table Data 5 Info													
(Read Only)													
	customer_key	version	date_from	date_to	customer_id	first_name	last_name	date_of_birth	city	state			
	1	1	(NULL)	(NULL)	(NULL)	OK	(NULL)	OK	(NULL)	OK	(NULL)	OK	
	2	1	1900-01-01 00:00:00	2015-01-01 00:00:00	C1050	5B	Ann	3B	Benson	6B	1993-01-03 00:00:00	Chicago	7B
	3	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C1197	5B	Panpan	6B	Bressler	8B	1992-01-30 00:00:00	PHIL...	12B
	4	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C16400	6B	Tairan	6B	Adelson	7B	1993-06-30 00:00:00	Pitt...	10B
	5	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C16474	6B	Marco	5B	Hussie	6B	1971-01-29 00:00:00	Pitt...	10B
	6	2	2015-01-01 00:00:00	2017-08-01 00:00:00	C1050	5B	BethAnn	7B	Benson	6B	1993-01-03 00:00:00	Chicago	7B
	7	3	2017-08-01 00:00:00	9999-01-01 00:00:00	C1050	5B	Ann	3B	Benson	6B	1993-01-03 00:00:00	Chicago	7B

First name is the type 2 variable, it inserts a new row with the first name changed to Ann. Rest of the columns were punch through SCD type 1 variables, therefore the details were overwritten.

	A	B	C	D	E	F	G	H	I	J
1	customer_id	first_name	last_name	date_of_birth	city	state	effective_date			
2	C1050	BethAnn	Cox	1/3/1993	New York	New York	1/1/2015			
3	C1050	Ann	Benson	1/3/1993	Chicago	Illinois	8/1/2017			
4	C1197	Panpan	Bressler	1/30/1992	PHILADELPHIA	Pennsylvania	1/1/2015			
5	C16400	Tairan	Adelson	6/30/1993	Pittsburgh	Pennsylvania	1/1/2015			
6	C16474	Marco	Hussie	1/29/1971	Pittsburgh	Pennsylvania	1/1/2015			
7										
8										

AutoSave OFF QueryOutput Search Nikhitha Kamath NK										
File Home Insert Draw Page Layout Formulas Data Review View Help										
K13										
	A	B	C	D	E	F	G	H	I	J
1	customer_key	version	date_from	date_to	customer_id	first_name	last_name	date_of_birth	city	state
2		1								
3		2	1	1900-01-01 00:00:00	2015-01-01 00:00:00	C1050	Ann	Benson	1993-01-03 00:00:00	Chicago
4		3	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C1197	Panpan	Bressler	1992-01-30 00:00:00	PHILADELPHIA
5		4	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C16400	Tairan	Adelson	1993-06-30 00:00:00	Pittsburgh
6		5	1	1900-01-01 00:00:00	4000-01-01 00:00:00	C16474	Marco	Hussie	1971-01-29 00:00:00	Pittsburgh
7		6	2	2015-01-01 00:00:00	2017-08-01 00:00:00	C1050	BethAnn	Benson	1993-01-03 00:00:00	Chicago
8		7	3	2017-08-01 00:00:00	9999-01-01 00:00:00	C1050	Ann	Benson	1993-01-03 00:00:00	Chicago