

Data Warehousing Concept Using ETL Process for SCD Type-2

K.Srikanth¹, N.V.E.S.Murthy², J.Anitha³

¹(Computer Science and Systems Engineering, Andhra University, India)

²(Computer Science and Systems Engineering, Andhra University, India)

³(Computer Science and Systems Engineering, Andhra University, India)

Abstract: SCD type 2 will store the entire history in the dimension table. In SCD type 2 effective date, the dimension table will have Start_Date and End_Date as the fields. If the End_Date is Null, then it indicates the current row. Know more about SCDs at Slowly Changing Dimensions Concepts. The new incoming record (changed/modified data set) replaces the existing old record in target. We will see how to implement the SCD Type 2 Effective Date in informatica. If there are retrospective changes made to the contents of the dimension, or if new attributes are added to the dimension which have different effective dates from those already defined, then this can result in the existing transactions needing to be updated to reflect the new situation. As an example consider the Employee dimension.

Keywords: ETL; Metadata; Mapping; Transformation.

I. INTRODUCTION

The beauty of this approach is it will maintain two versions, you will find two records the older version and the current version. In other words it maintains history. The thing to be noticed here is if there is any update in the salary of any employee then the history of that employee is displayed with the current date as the start date and the previous date as the end date. As in case of any SCD Type 2 implementation[1], here we need to first find out the set of SCD2 records which qualify for either INSERT or INSERT/UPDATE. Based on this approach, a typical mapping will contain expression, router and update strategy transformations but will not contain any lookup transformation.

Again we can implement Type 2 in following methods

1. Versioning
2. Effective Dates
3. By setting Current Flag values/Record Indicators.
4. We will divide the steps to implement the SCD type 2 Effective Date mapping into four parts.

II. SCD TYPE 2 EFFECTIVE DATE IMPLEMENTATION

Here we will see the basic set up and mapping flow require for SCD type 2 Effective Date. The steps involved are

Implementation:

Source:

```
SQL> SELECT * FROM EMP;
```

EMPNO	ENAME	JOB	MGR	HIREDATE	SAL	COMM	DEPTNO
7369	SMITH	CLERK	7902	17-DEC-80	2300		20
7499	ALLEN	SALESMAN	7698	20-FEB-81	1600	300	30
7521	WARD	SALESMAN	7698	22-FEB-81	1250	500	30
7566	JONES	MANAGER	7839	02-APR-81	2975		20
7654	MARTIN	SALESMAN	7698	28-SEP-81	1250	1400	30
7698	BLAKE	MANAGER	7839	01-MAY-81	2850		30
7782	CLARK	MANAGER	7839	09-JUN-81	2450		10
7788	SCOTT	ANALYST	7566	19-APR-87	3000		20
7839	KING	PRESIDENT		17-NOV-81	5200		10
7844	TURNER	SALESMAN	7698	08-SEP-81	1500	0	30
7876	ADAMS	CLERK	7788	23-MAY-87	1100		20
7900	JAMES	CLERK	7698	03-DEC-81	950		30
7902	FORD	ANALYST	7566	03-DEC-81	3000		20
7934	MILLER	CLERK	7782	23-JAN-82	1300		10

14 rows selected.

Table 1: Oracle SQL Query On EMP Table

- Create the source and dimension tables in the database using Table 1.
- Open the mapping designer tool, source analyzer and either create or import the source definition.
- Go to the Warehouse designer or Target designer and import the target definition[2].
- Go to the mapping designer tab and create new mapping.
- Drag the source into the mapping[7].
- Go to the toolbar, Transformation and then Create.
- Select the lookup Transformation, Figure 1. enter a name and click on create. You will get a window as shown in the below image.

Figure 1: Creating Lookup Transformation ports logic

- Edit the lookup transformation, go to the ports tab and remove unnecessary ports. Just keep only Empkey, EMPNO and location ports in the lookup transformation. Create a new port (new_flag, update_flag) in the lookup transformation[3]. This new port needs to be connected to the output port of the Expression transformation.
- Go to the conditions tab of the lookup transformation and enter the condition as EMPNO= EMPNO1.
- Go to the properties tab of the LKP transformation and enter the below query in Lookup SQL Override[1]. Alternatively you can generate the SQL query by connecting the database in the Lookup SQL Override expression editor and then add the WHERE clause.

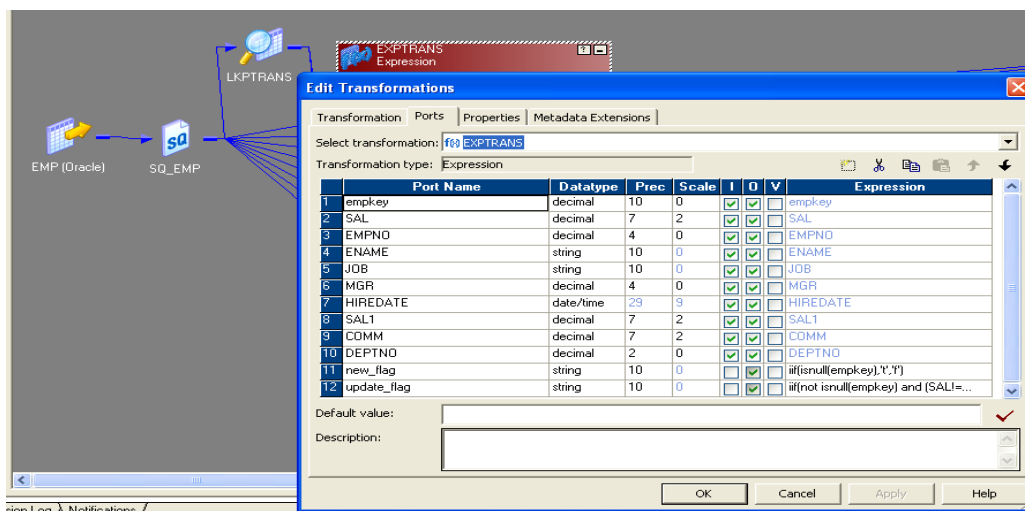


Figure 2: Creating Expression Transformation ports logic

You can add ports to expression transformation either by selecting and dragging ports from other transformations or by opening the expression transformation and create ports manually[4], Figure 2. We can add the port new_flag and update_flag using string datatype. In expression transformation implement the employee key either true or false.

1. IIF(ISNULL(EMPKEY),'T','F');
2. IIF(NOT ISNULL(EMPKEY) AND (SAL!=SAL1),'T','F');

II. SCD TYPE 2 EFFECTIVE DATE IMPLEMENTATION

In this part, we will identify the new records and insert them into the target with Begin Date as the current date. The steps involved are:

- Go the properties tab of filter transformation and enter the filter condition as New_Flag=T and Update_Flag=T[5].
- Edit Router Transformation select groups port writing the Group Filter Condition in the inset and update flags, Figure 3.

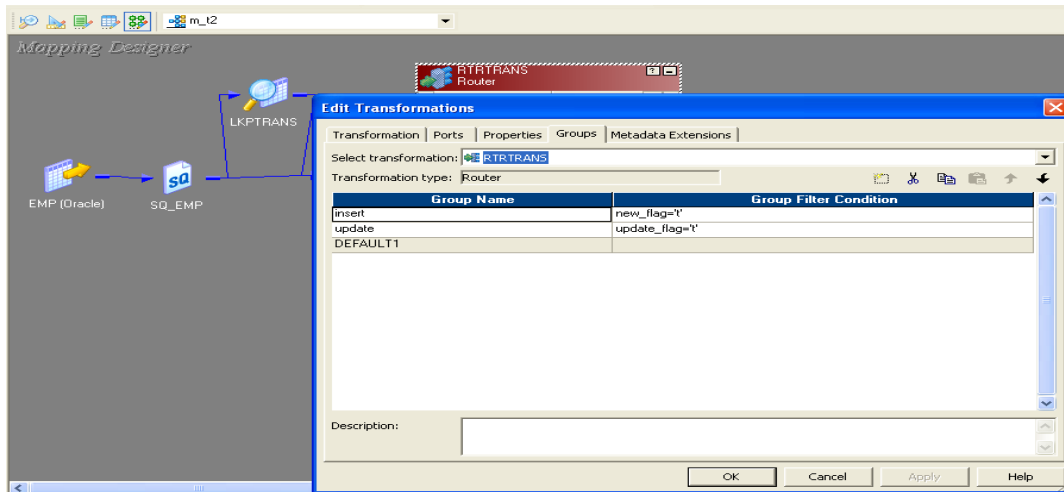


Figure 3: Creating Router Transformation Groups logic
 Insert : New_Flag='T' Update: Update_Flag='T'

III. SCD TYPE 2 EFFECTIVE DATE IMPLEMENTATION

In this part, we will identify the changed records and insert them into the target with Begin Date as the current date. Figure 4, The steps involved are:

- Now connect the ports of expression transformation (Nextval, Start_Date) to the Target definition ports (Emp_Key, End_Date)[6]. The part of the mapping flow is shown in the below image.
- Now drag the target definition into the mapping and connect the appropriate ports of update strategy transformation to the target definition.
- Drag and connect the NextVal port of sequence generator to the Expression transformation. In the expression transformation create a new output port (Start_Date and End_Date) and assign value SYSDATE to it.

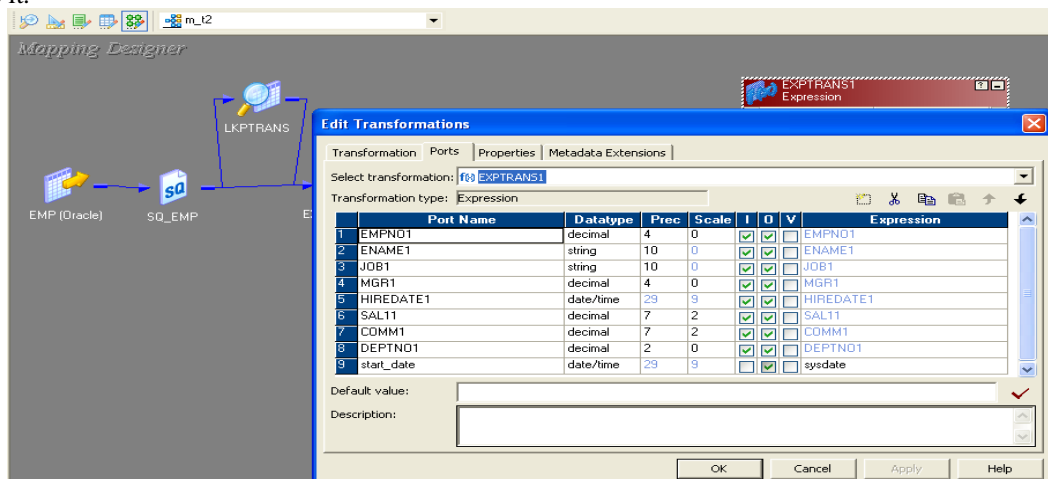


Figure 4: Creating Expression Transformation ports logic
 Start_date: Sysdate

IV. SCD TYPE 2 EFFECTIVE DATE IMPLEMENTATION

In this part, we will update the changed records in the dimension table with End Date as current date.

- Go to the ports tab of expression transformation and create a new output port (Start_Date and End_Date with date/time data type). Assign a value SYSDATE to this port[5].
- Now create an update strategy transformation and drag the ports of the expression transformation into it. Go to the properties tab and enter the update strategy expression as DD_UPDATE.
- Drag the target definition into the mapping and connect the appropriate ports of update strategy to it. Figure 5, The complete mapping image is shown below.

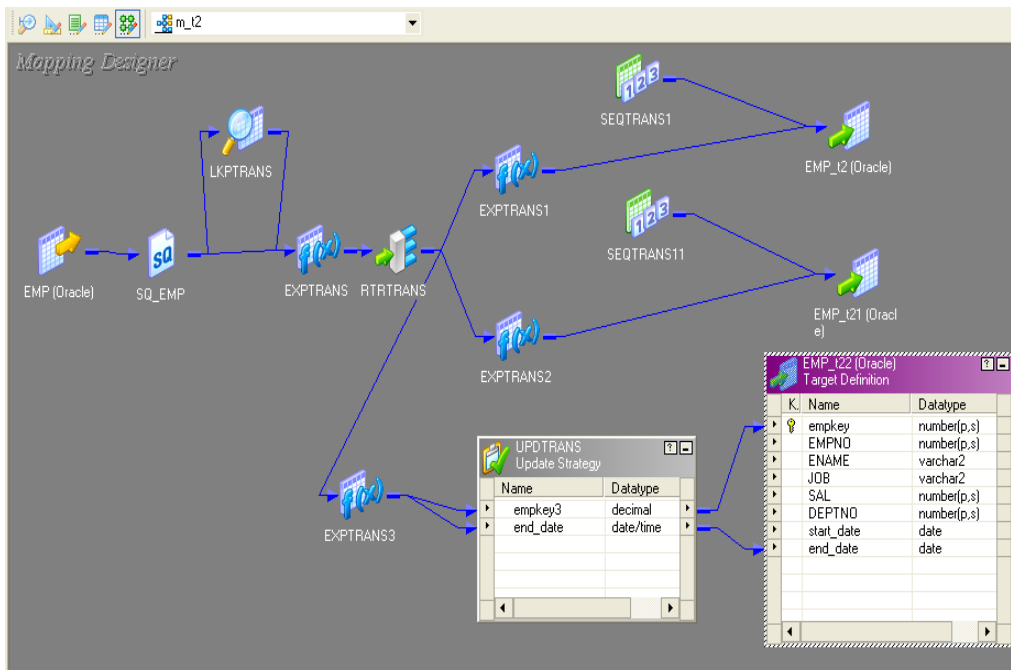


Figure 5: Update Strategy Transformation using End-Date
End_Date: Sysdate

The complete Slowly Changing Dimension Mapping Design flow, Figure 6. This flow will provide completion information of SCD-Type-2 source data how to load target, maintain the data processing.

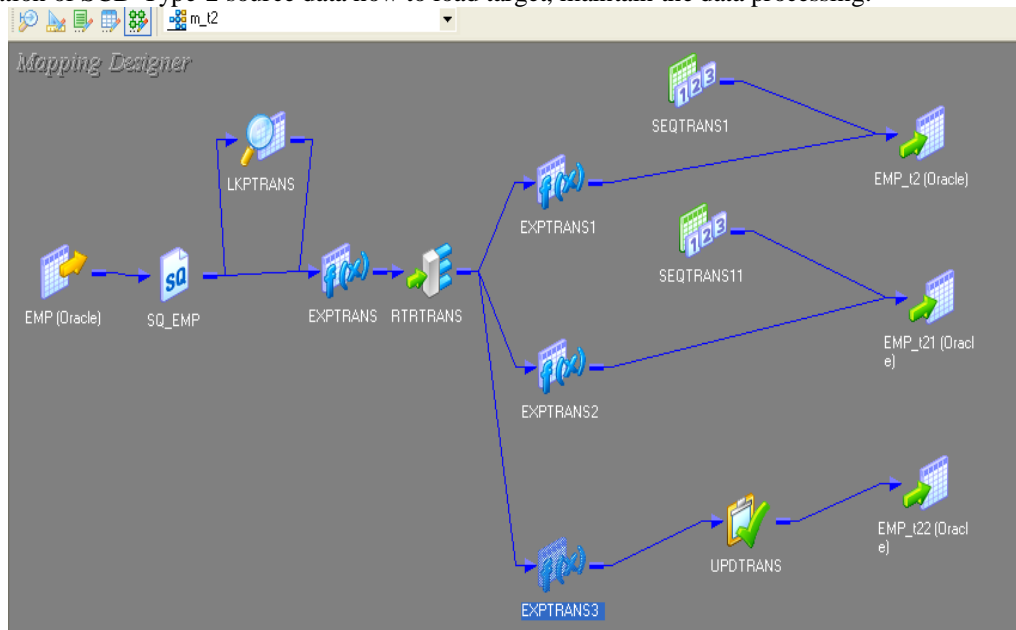


Figure 6: Slowly Changing Dimensions (SCDs) Flow

V. RESULTS

```
SQL> select * from emp_t2;
```

EMPKEY	EMPNO	ENAME	JOB	SAL	DEPTNO	START_DAT	END_DATE
1	7369	SMITH	CLERK	2300	20	02-FEB-12	
2	7499	ALLEN	SALESMAN	1600	30	02-FEB-12	
3	7521	WARD	SALESMAN	1250	30	02-FEB-12	
4	7566	JONES	MANAGER	2975	20	02-FEB-12	
5	7654	MARTIN	SALESMAN	1250	30	02-FEB-12	
6	7698	BLAKE	MANAGER	2850	30	02-FEB-12	
7	7782	CLARK	MANAGER	2450	10	02-FEB-12	
8	7788	SCOTT	ANALYST	3000	20	02-FEB-12	
9	7839	KING	PRESIDENT	5200	10	02-FEB-12	
10	7844	TURNER	SALESMAN	1500	30	02-FEB-12	
11	7876	ADAMS	CLERK	1100	20	02-FEB-12	
12	7900	JAMES	CLERK	950	30	02-FEB-12	
13	7902	FORD	ANALYST	3000	20	02-FEB-12	
14	7934	MILLER	CLERK	1300	10	02-FEB-12	
15	1111	SRIKANTH	PHD	9000	40	02-FEB-12	

15 rows selected.

Table 2: Oracle SQL Query On EMP Table Target Data

```
SQL> insert into emp values(2345,'dileep','dfd',2346,'23-feb-1986',467,455,40);
1 row created.
```

Once load the target data after write oracle queries in insert data and update the values Using connect the employee table. Table 2, Table 3. Below oracle table insert, update display the new type 2 complete updated data.

```
SQL> select * from emp_t2;
```

EMPKEY	EMPNO	ENAME	JOB	SAL	DEPTNO	START_DAT	END_DATE
207	7369	SMITH	CLERK	2300	20	02-FEB-12	
208	7499	ALLEN	SALESMAN	1600	30	02-FEB-12	
209	7521	WARD	SALESMAN	1250	30	02-FEB-12	
210	7566	JONES	MANAGER	2975	20	02-FEB-12	
211	7654	MARTIN	SALESMAN	1250	30	02-FEB-12	
212	7698	BLAKE	MANAGER	2850	30	02-FEB-12	
213	7782	CLARK	MANAGER	2450	10	02-FEB-12	
214	7788	SCOTT	ANALYST	3000	20	02-FEB-12	
215	7839	KING	PRESIDENT	5200	10	02-FEB-12	
216	7844	TURNER	SALESMAN	1500	30	02-FEB-12	
217	7876	ADAMS	CLERK	1100	20	02-FEB-12	
218	7900	JAMES	CLERK	950	30	02-FEB-12	
219	7902	FORD	ANALYST	3000	20	02-FEB-12	
220	7934	MILLER	CLERK	1300	10	02-FEB-12	
221	1111	SRIKANTH	PHD	2000	40	02-FEB-12	
222	2345	dileep	dfd	1200	40	02-FEB-12	02-FEB-12
223	2345	dileep	dfd	1000	40	02-FEB-12	

17 rows selected.

Table 3: Oracle SQL Query On EMP Table Updated Target Data

Source Data: Table 1
 Target Data : Table 2
 Updated Target Data: Table 3

VI. CONCLUSIONS

Extraction-Transformation-Loading (ETL) tools are pieces of software responsible for the extraction of data from several sources. In this paper, we have focused on the problem A Type One change updates only the attribute, doesn't insert new records, and affects no keys. It is easy to implement but does not maintain any history of prior attribute values. **Slowly Changing Dimensions (SCDs)** are dimensions that have data that changes slowly, rather than changing on a time-based, regular schedule. In SCD type 2 effective date, the dimension table will have Start_Date and End_Date as the fields. If the End_Date is Null, then it indicates the current row. Know more about SCDs at Slowly Changing Dimensions Concepts. The new incoming record

(changed/modified data set) replaces the existing old record in target. Comprehensive ETL criteria were identified. testing procedures were developed. and this work was applied to commercial ETL tools. The study covered all major aspects of ETL usage and can be used to effectively compare and evaluate various ETL tools. We can implement on SCD TYPE-2 based on SCD TYPE-1 and new fields like Versioning, Effective Dates, By setting Current Flag values/Record Indicators.

REFERENCES

Journal Papers:

- [1] I. William, S. Derek, and N. Genia, DW 2.0: The Architecture for the Next Generation of Data Warehousing. Burlington, MA: Morgan Kaufman, 2008, pp. 215-229.
- [2] R. J. Davenport, September 2007. [Online] ETL vs. ELT: A Subjective View. In Source IT Consulting Ltd., U.K. Available at: http://www.insource.co.uk/pdf/ETL_ELT.pdf.
- [3] T. Jun, C. Kai, Feng Yu, T. Gang, "The Research and Application of ETL Tools in Business Intelligence Project," in Proc. International Forum on Information Technology and Applications, 2009, IEEE, pp.620-623.

Books:

- [4] Kimball, R., Caserta, J.: The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, loading, Conforming, and Delivering Data. John Wiley & Sons, 2004.
- [5] Informatica Power Center, Available at www.informatica.com/products/data_integration/power_center/default.htm

Theses:

- [6] ALKIS SIMITSIS , Dipl. Electrical and Computer Engineering (2000). Modeling and Optimization of Extraction-Transformation-Loading (ETL) Processes in Data Warehouse Environments

Proceedings Papers:

- [7] K.Srikanth, N.V.S.Murthy, J.Anitha : "Data Warehousing Concept Using ETL Process For SCD Type-1" Conf. on TIJCSA , Volume 1, No. 10, December 2012 ISSN – 2278-1080.