ABSTRACT

In a data warehousing system, Change Data capture (CDC) plays an important role not just in making the data warehouse (DWH) aware of the change but also providing a means of flowing the change to the DWH marts and reporting tables so that we see the current and latest version of the truth. This, together with Slowly Changing Dimensions (SCD), creates a cycle which runs the DWH and provides valuable insights in the history and decision making future when used by machine learning, pattern recognition and forecasting methods. What if the source has no CDC? It would be an ETL nightmare to identify the exact change and report the absolute truth. This hybrid approach shows how to increase the data and metadata capture and quality in data warehouses. Application of this hybrid approach can provide information that is otherwise hidden in areas such as Audit and Fraud detection, while the efficiency provides opportunities for rapidly changing dimensions.

INTRODUCTION

Often it is seen that these two processes are separated and applied at different stages of the DWH, e.g. CDC is applied at the source area like the DBMS which will give you the changed records in the Staging area (often a stored process or a trigger event applied on the schema will do it) whereas the SCD is done in the Transform area where we act on the changed records to process them as per the business rules. If these two different processes can be combined in just a single process where just one single transform does both jobs of “identifying the change” and “applying the change to the DWH” then we can save significant processing times and value resources of the system doing all in one pass. Hence, we came up with a Hybrid SCD with CDC approach for this.

Figure 1. Hybrid transform to process CDC and SCD Type2 for sources WITHOUT CDC
Figure 2. Configuration Options for the Hybrid Transform

1. **Business Key** columns *(required)* – Select the Business Keys (BK) used to compare the generate Surrogate Keys.

2. **Surrogate Key** column *(required)* – The column in which Surrogate Key (SK) values will be populated.

3. **Current Record Indicator** column *(required)* – The column will have a value of 0/1. If 0 then record is *old* and 1 if the record is *current*.

4. **Business Start Date** column *(optional)* – This is an optional selection. There are scenarios where we have a sources which might have a business start date-time.

5. **Change Data Capture (CDC)** columns *(optional)* – In our design we have kept it as optional so that if a source is changing then we will have to consider all other columns except BKs and SKs as changing columns over time.

**USE CASES**

We will consider three use cases which were solved by using our hybrid approach,

1. **Sources WITH** Effective Dates – Often CDC is dependent on the last updated date of the record in the source system. This together with Business Start and End dates can identify the changed records or just the update date will suffice.

2. **Sources WITHOUT** Effective Dates – There is virtually no scope of identifying the changed records as the source system is constantly refreshing the changes instead of inserting a new record for the change. This would be like starting off with a clean slate before any change is made. The changed data is lost completely unless we have a DWH to look for a change here.

3. **Source records that DISAPPEAR** whether by soft/logical delete or by physical delete. Records are sometimes removed from data sources and such losses may need to be identified as such rather than being end-dated.
THE DIFFERENTIATOR – HYBRID PROCESS

Source table WITHOUT CDC

DIMENSION table exists?

Create DIMENSION table

NO

Compare CDC records

CDC tables exist?

Create CDC tables

NO

Identify NEW / UPDATED records

YES

NEW / UPDATED records existed?

Add records to CDC tables

HYBRID PROCESS

NO

End Process

NO

Identify and add MISSING records to Control Table

Insert NEW / UPDATED records to DIMENSION table

Figure 3. The Differentiator – Hybrid process Flow
1. SOURCE TABLE WITHOUT CDC

This is the source table which doesn’t have CDC enabled. This can be any SAS dataset or any other Database where they have failed to enable CDC on. I have taken a sample of employee’s dataset of the Chicago police which is available online.

![Figure 4. Source table without CDC](image)

2. CREATE DIMENSION TABLE

For the first time we prepare the Dimension table. There are no SKs and Generated Keys (Business Start and End dates, Current record indicator and Load Date-time) and so the process generates them.

![Figure 5. First run of SCD Type2 - SKs and Generated Keys](image)

3. IDENTIFY THE NEW AND UPDATED RECORDS

For all the subsequent runs, the hybrid process runs by identifying the newly inserted records (NEW) and the records which are updated (UPDATED). We compared the Dimension table with the Source table, finally a column CDC_RUN_DTTM will identify when the updated or new records were inserted.

![Figure 6. Identify the NEW and UPDATED records from the source](image)
4. **COMPARE CDC RECORDS**

This step reduces the overhead of processing the CDC multiple times at any point in time. The flow of steps below avoid processing the CDC multiple times by ending the process if the change records already exist as *update or new* records.

![Diagram](image-url)

**Figure 7. Daily CDC comparison process**

Let's say, the *updated/new* records exist today, then there are two scenarios

- **No CDC tables exist** – This is the second time the SCD type 2 process has run and there are no CDC tables then create them.

- **CDC tables exist** – For the 3rd and subsequent runs compare the *CDC update and CDC new* tables of today to that of previous day. If there is a change then add to DIM table else exit the process.

The above design eliminates the overhead of additional CPU and IO processing times by *NOT processing previously processed UPDATED and NEW records* over again.

5. **ADD NEW/UPDATED RECORDS TO THE DIMENSION TABLE**

Once all the *new* and *updated* records have been identified then add them to the Dimension table as below,

For *updated* records

1. Close the Business End Date for the *old record* – set it to 31JAN9999 (future date/ NULL will do)
2. Update the Current Record Indicator of the *updated record* to ‘1’ and set the *old record* to ‘0’

For *new* records – *INSERT* the new records into dimension table with Current record indicator equal to ‘1’ Business Start Date as today’s date and End date to a future date.
DISAPPEARING RECORDS – ROGUE RECORDS

There are scenarios where a record is soft/hard deleted from the source system. This might happen either accidentally or on purpose (sometime to hide audit trails in money laundering). Our design caters to this and provides information about missing/deleted records in the system to answer the questions,

- What record was deleted from the system?
- When the record last seen in the system?

This will be valuable in businesses where we need audit trail of the records being processed, e.g. Anti-Money Laundering (AML), Electronic Record keeping systems.

PERFORMANCE

The below chart shows the performance comparison of the SAS® Data Integration Studio SCD Type 2 transform against the Hybrid. It is clear that there is a large improvement in performance with the hybrid transform as compared to the regular SCD Type 2, in addition to get better audit trail features.
CONCLUSION

This approach strengthens audit level capabilities for data warehousing. The efficiency means multiple changes within one day can occur to increase the visibility of changes. **Without such an approach information and transactions are lost before the data warehouse can read them.** Insurance, Audit, Fraud detection and Intelligence industries will find great benefit in applying this mechanism to capture transactions.

It is these low frequency high value audit trails that demand the detail, precision and completeness provided by this approach.

COPYRIGHT

Compare CDC records® This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of Vishant Bhat, email: vishant.bhat@gmail.com.

REFERENCES

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Vishant Bhat
SAS Consultant
Email: vishant.bhat@gmail.com

Tony Blanch
SAS Consultant
Email: abdata4@gmail.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.