Paper 7560-2016

Processing CDC and SCD type2 for sources without CDC – Hybrid approach

Vishant Bhat, SAS Consultant; Tony Blanch, SAS Consultant

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) create 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.

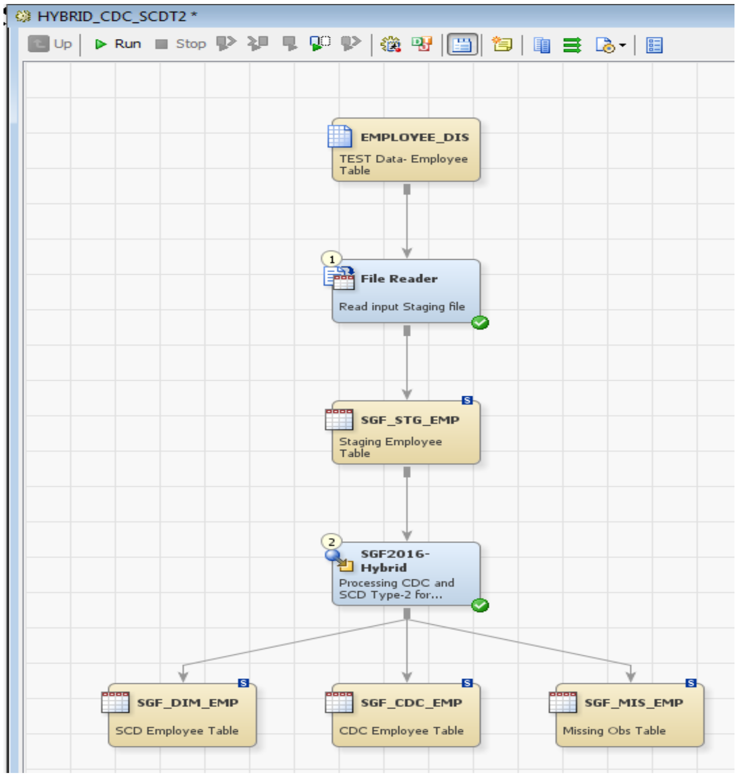
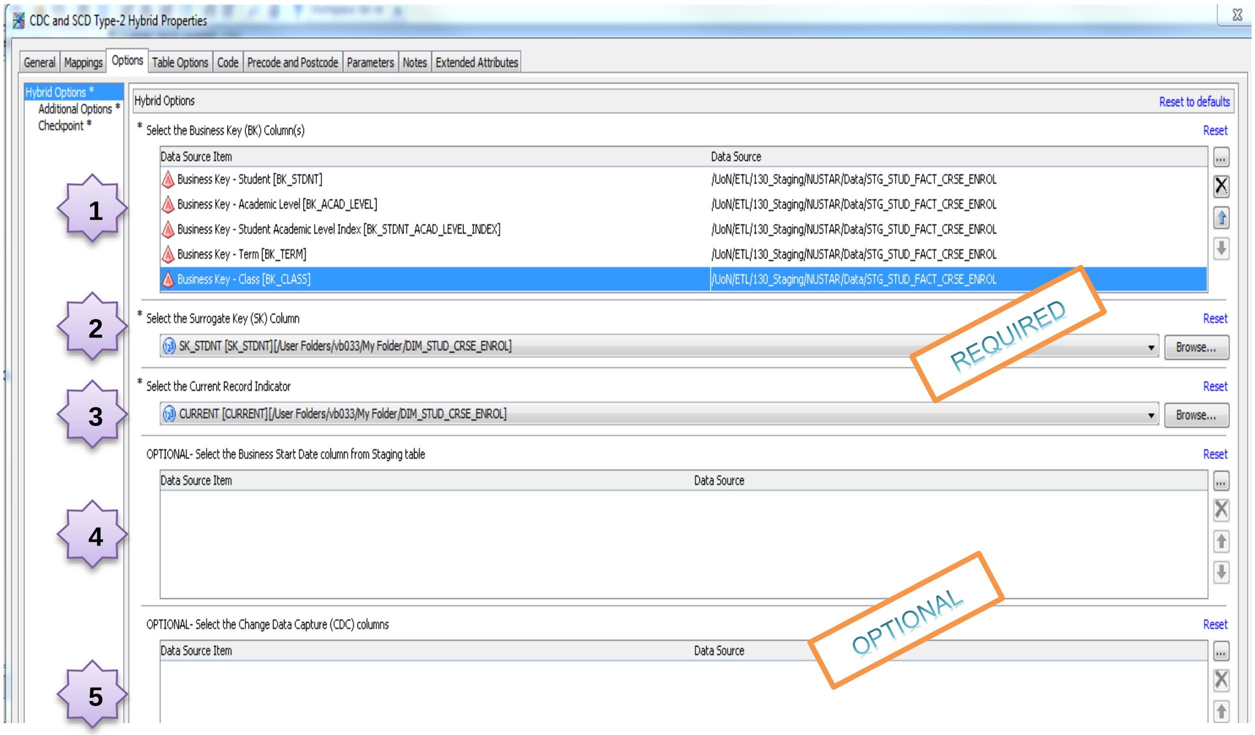


Figure 1. Hybrid transform to process CDC and SCD Type2 for sources WITHOUT CDC



Hybrid Options

* Select the Business Key (BK) Column(s)

Data Source Item	Data Source
Business Key - Student [BK_STDNT]	/UoN/ETL/130_Staging/NUSTAR/Data/STG_STUD_FACT_CRSE_ENROL
Business Key - Academic Level [BK_ACAD_LEVEL]	/UoN/ETL/130_Staging/NUSTAR/Data/STG_STUD_FACT_CRSE_ENROL
Business Key - Student Academic Level Index [BK_STDNT_ACAD_LEVEL_INDEX]	/UoN/ETL/130_Staging/NUSTAR/Data/STG_STUD_FACT_CRSE_ENROL
Business Key - Term [BK_TERM]	/UoN/ETL/130_Staging/NUSTAR/Data/STG_STUD_FACT_CRSE_ENROL
Business Key - Class [BK_CLASS]	/UoN/ETL/130_Staging/NUSTAR/Data/STG_STUD_FACT_CRSE_ENROL

* Select the Surrogate Key (SK) Column

SK_STDNT [SK_STDNT][/User Folders/vb033/My Folder/DIM_STUD_CRSE_ENROL]

* Select the Current Record Indicator

CURRENT [CURRENT][/User Folders/vb033/My Folder/DIM_STUD_CRSE_ENROL]

OPTIONAL- Select the Business Start Date column from Staging table

Data Source Item	Data Source

OPTIONAL- Select the Change Data Capture (CDC) columns

Data Source Item	Data Source

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

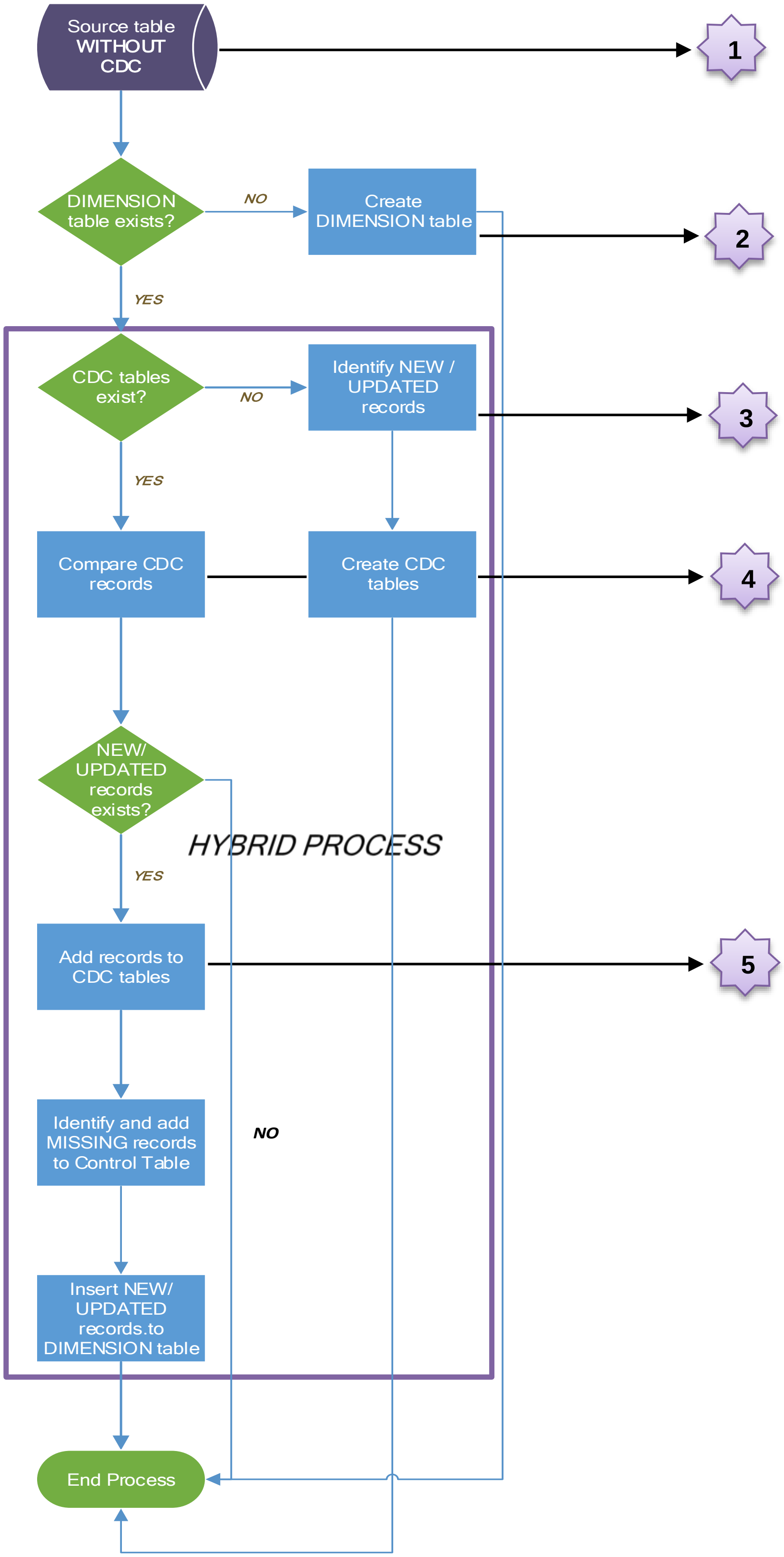
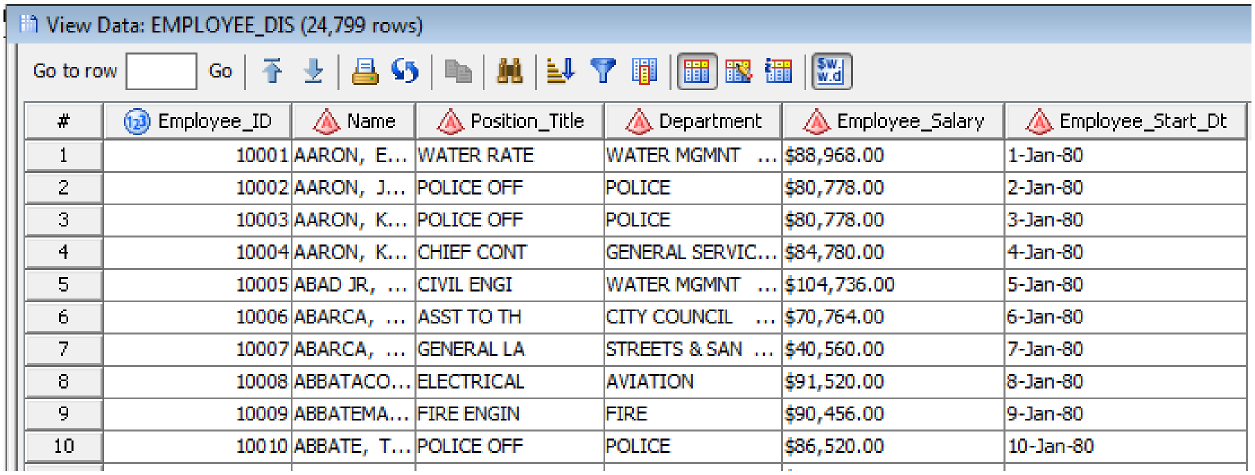

Source table
WITHOUT
CDC
1
DIMENSION table exists?
NO
Create DIMENSION table
2
YES
CDC tables exist?
NO
Identify NEW / UPDATED records
3
YES
Compare CDC records
Create CDC tables
4
NEW/ UPDATED records exists?
HYBRID PROCESS
YES
Add records to CDC tables
5
Identify and add MISSING records to Control Table
NO
Insert NEW/ UPDATED records.to DIMENSION table
End Process

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.



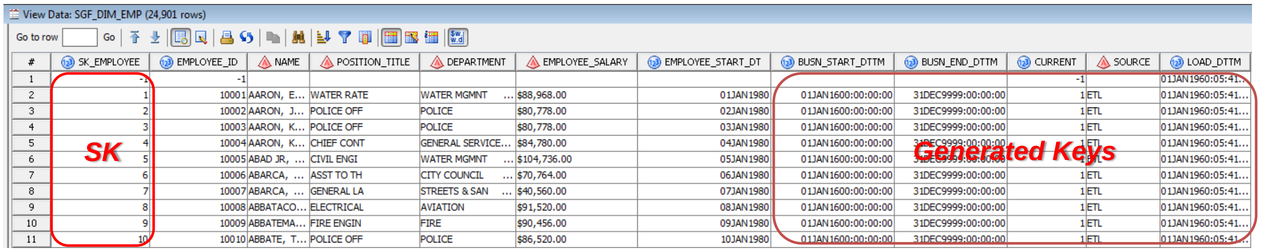
View Data: EMPLOYEE_DIS (24,799 rows)

#	Employee_ID	Name	Position_Title	Department	Employee_Salary	Employee_Start_Dt
1	10001	AARON, E...	WATER RATE	WATER MGMNT ...	$88,968.00	1-Jan-80
2	10002	AARON, J...	POLICE OFF	POLICE	$80,778.00	2-Jan-80
3	10003	AARON, K...	POLICE OFF	POLICE	$80,778.00	3-Jan-80
4	10004	AARON, K...	CHIEF CONT	GENERAL SERVIC...	$84,780.00	4-Jan-80
5	10005	ABAD JR, ...	CIVIL ENGI	WATER MGMNT ...	$104,736.00	5-Jan-80
6	10006	ABARCA, ...	ASST TO TH	CITY COUNCIL ...	$70,764.00	6-Jan-80
7	10007	ABARCA, ...	GENERAL LA	STREETS & SAN ...	$40,560.00	7-Jan-80
8	10008	ABBATACO...	ELECTRICAL	AVIATION	$91,520.00	8-Jan-80
9	10009	ABBATEMA...	FIRE ENGIN	FIRE	$90,456.00	9-Jan-80
10	10010	ABBATE, T...	POLICE OFF	POLICE	$86,520.00	10-Jan-80

Figure 4. Source table without CDC

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.



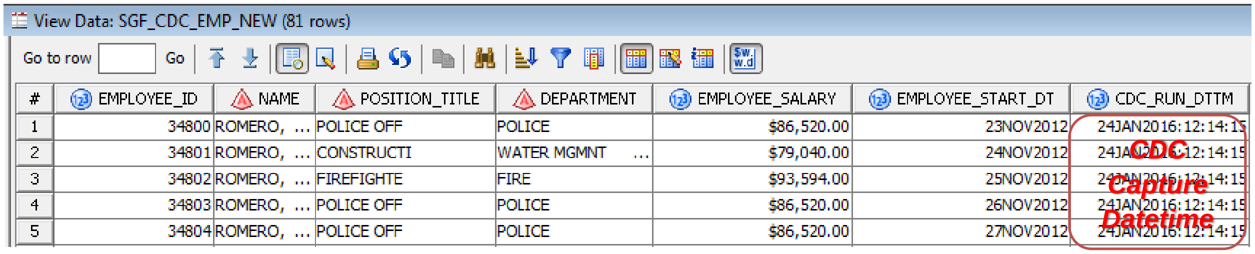
View Data: SGF_DIM_EMP (24,901 rows)

#	SK_EMPLOYEE	EMPLOYEE_ID	NAME	POSITION_TITLE	DEPARTMENT	EMPLOYEE_SALARY	EMPLOYEE_START_DT	BUSN_START_DTTM	BUSN_END_DTTM	CURRENT	SOURCE	LOAD_DTTM
1	-1	-1								-1		01JAN1960:05:41...
2	1	10001	AARON, E...	WATER RATE	WATER MGMNT ...	$88,968.00	01JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...
3	2	10002	AARON, J...	POLICE OFF	POLICE	$80,778.00	02JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...
4	3	10003	AARON, K...	POLICE OFF	POLICE	$80,778.00	03JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...
5	4	10004	AARON, K...	CHIEF CONT	GENERAL SERVICE...	$84,780.00	04JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...
6	5	10005	ABAD JR, ...	CIVIL ENGI	WATER MGMNT ...	$104,736.00	05JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...
7	6	10006	ABARCA, ...	ASST TO TH	CITY COUNCIL ...	$70,764.00	06JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...
8	7	10007	ABARCA, ...	GENERAL LA	STREETS & SAN ...	$40,560.00	07JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...
9	8	10008	ABBATACO...	ELECTRICAL	AVIATION	$91,520.00	08JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...
10	9	10009	ABBATEMA...	FIRE ENGIN	FIRE	$90,456.00	09JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...
11	10	10010	ABBATE, T...	POLICE OFF	POLICE	$86,520.00	10JAN1980	01JAN1600:00:00:00	31DEC9999:00:00:00	1	ETL	01JAN1960:05:41...

Figure 5. First run of SCD Type2 - SKs and Generated Keys

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.



View Data: SGF_CDC_EMP_NEW (81 rows)

#	EMPLOYEE_ID	NAME	POSITION_TITLE	DEPARTMENT	EMPLOYEE_SALARY	EMPLOYEE_START_DT	CDC_RUN_DTTM
1	34800	ROMERO, ...	POLICE OFF	POLICE	$86,520.00	23NOV2012	24JAN2016:12:14:15
2	34801	ROMERO, ...	CONSTRUCTI	WATER MGMNT ...	$79,040.00	24NOV2012	24JAN2016:12:14:15
3	34802	ROMERO, ...	FIREFIGHTE	FIRE	$93,594.00	25NOV2012	24JAN2016:12:14:15
4	34803	ROMERO, ...	POLICE OFF	POLICE	$86,520.00	26NOV2012	24JAN2016:12:14:15
5	34804	ROMERO, ...	POLICE OFF	POLICE	$86,520.00	27NOV2012	24JAN2016:12:14:15

Figure 6. Identify the NEW and UPDATED records from the source

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.

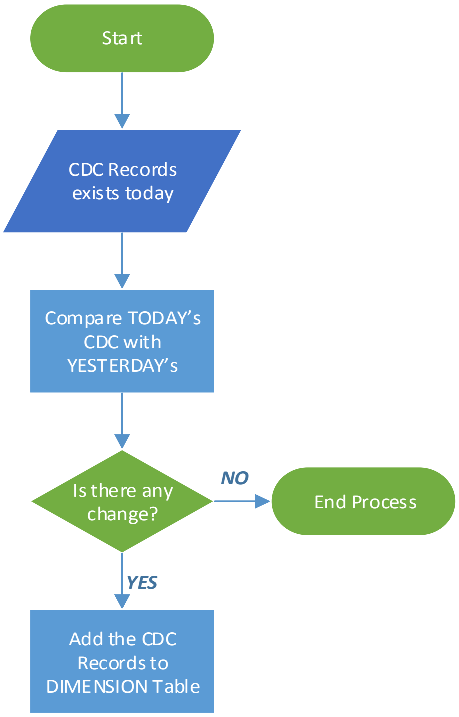

Start
CDC Records exists today
Compare TODAY's CDC with YESTERDAY's
Is there any change?
NO
End Process
YES
Add the CDC Records to DIMENSION Table

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.

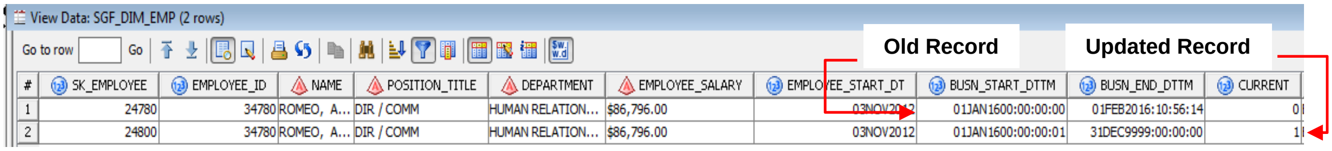
5. ADD NEW/UPDATED RECORDS TO THE DIMENSION TABLE

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.



View Data: SGF_DIM_EMP (2 rows)

#	SK_EMPLOYEE	EMPLOYEE_ID	NAME	POSITION_TITLE	DEPARTMENT	EMPLOYEE_SALARY	EMPLOYEE_START_DT	BUSN_START_DTTM	BUSN_END_DTTM	CURRENT
1	24780	34780	ROMEO, A...	DIR / COMM	HUMAN RELATION...	$86,796.00	03NOV2012	01JAN1600:00:00:00	01FEB2016:10:56:14	0
2	24800	34780	ROMEO, A...	DIR / COMM	HUMAN RELATION...	$86,796.00	03NOV2012	01JAN1600:00:00:01	31DEC9999:00:00:00	1

Figure 8. Updated Record in DIMENSION table

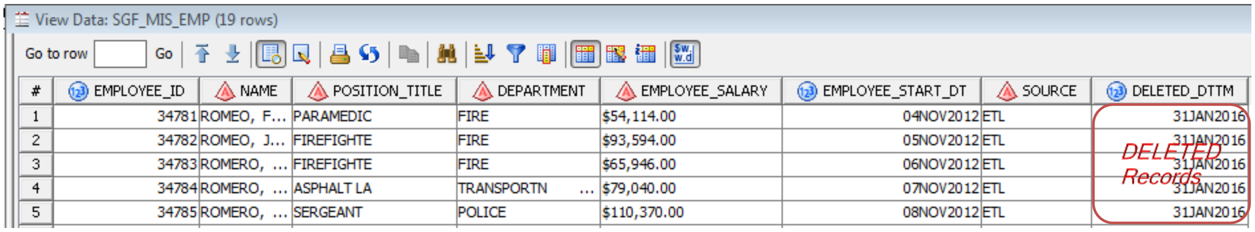
DISAPPEARING RECORDS – ROGUE RECORDS

There are scenarios where a record is ***soft/hard deleted from the source system.*** This might happen either accidently 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?

The ***DELETED_DTTM*** column gives us the date the record was deleted from 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***



View Data: SGF_MIS_EMP (19 rows)

#	EMPLOYEE_ID	NAME	POSITION_TITLE	DEPARTMENT	EMPLOYEE_SALARY	EMPLOYEE_START_DT	SOURCE	DELETED_DTTM
1	34781	ROMEO, F...	PARAMEDIC	FIRE	$54,114.00	04NOV2012	ETL	31JAN2016
2	34782	ROMEO, J...	FIREFIGHTE	FIRE	$93,594.00	05NOV2012	ETL	31JAN2016
3	34783	ROMERO, ...	FIREFIGHTE	FIRE	$65,946.00	06NOV2012	ETL	31JAN2016
4	34784	ROMERO, ...	ASPHALT LA	TRANSPORTN ...	$79,040.00	07NOV2012	ETL	31JAN2016
5	34785	ROMERO, ...	SERGEANT	POLICE	$110,370.00	08NOV2012	ETL	31JAN2016

Figure 9. MISSING records - Source records which are not present as of today

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.

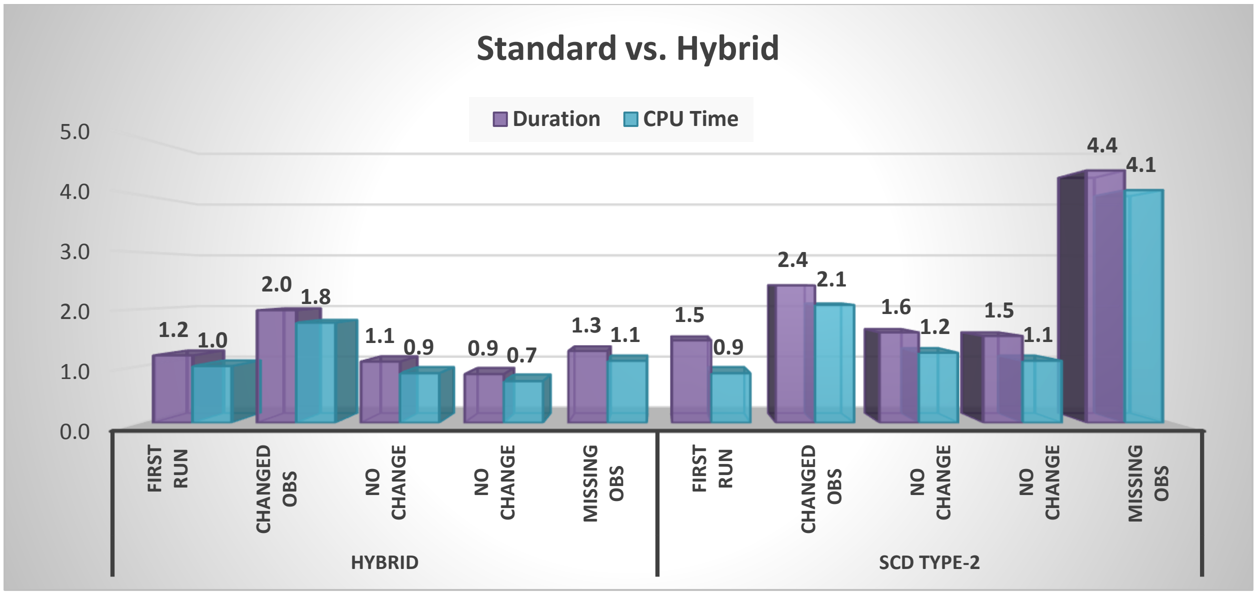


Figure 10. Standard vs. Hybrid comparison - Duration and CPU Time

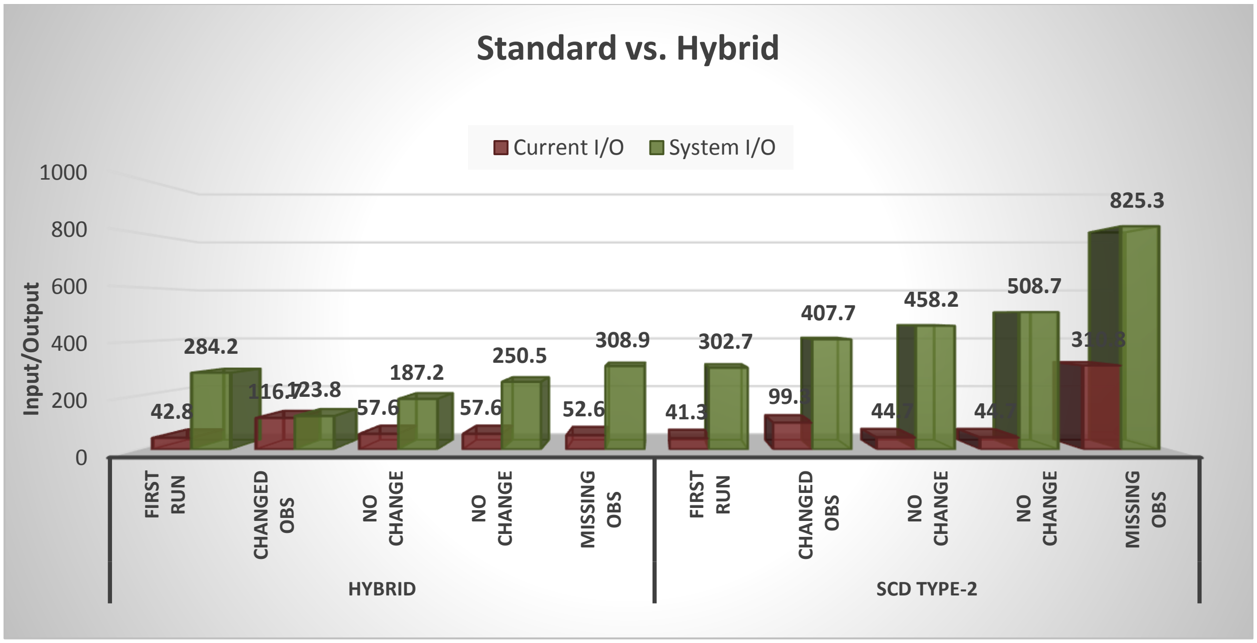


Figure 11. Standard vs. Hybrid comparison - Current and System IO

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

REFERENCES

- Sample data from Chicago Police department open dataset - https://data.cityofchicago.org/Administration-Finance/Current-Employee-Names-Salaries-and-Position-Title/xzkq-xp2w

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