1.0 Introduction

We have lots of different methods used in understanding and implementing Type 2 SCD in Snowflake and we will get lot of posts on this in different blogs. I am also trying to provide same topic but with different perspective.

I have tried to provide a generic approach using simple Snowflake stored procedures in implementing Type 2 SCD. The process is reusable and can be implemented with snowflake components easily with some configurations.

2.0 Components Used

- Tables (RAW/STG/Core/Config)
- Snowflake Streams
- Snowflake Stored Procedures

2.1 Type 2 SCD Components

In this process for performing SCD Type 2 for any table we need 3 sets of tables.

1. RAW Table – Load data from files in this table
2. STG Table – It holds the previous day data to match with current day data in the RAW table. The stream will be created in this table.
3. Core Table – This is the main table hold the data with Type 2 SCD.
4. Config Table – This configuration table holds the metadata details of the table like columns, keys, columns to update, columns to insert.

Snowflake Stored Procedures are used to capture the change between data and apply the changes. Following stored procedures are used.

1. **SP_CDC_MAIN** – Parameterized stored procedure is a wrapper script on top of SP_CHANGE_CAPTURE and SP_CHANGE_APPLY for execution of the merge statements created by these 2 stored procedures.

2. **SP_CHANGE_CAPTURE** - Parameterized stored procedure is capturing the changes between previous day data and current day data and create the merge statement to update the data in the STG table.

3. **SP_CHANGE_APPLY** - Parameterized stored procedure is applying the changes in the Core table through STG table and the Stream.

4. **SP_CHANGE_APPLY_DELETE** – Parameterized stored procedure applying soft deletes in transform table for full file processing in case record not in source but in target i.e. records deleted from source.

5. **SP_CHANGE_APPLY_DELETE_STG** – Parameterized stored procedure to delete the stage records due to the soft deletes in transform table for full file processing in case record not in source but in target i.e. records deleted from source.

6. **CRT_CDC_CONFIG** – Parameterized stored procedure to create CDC config table entries.

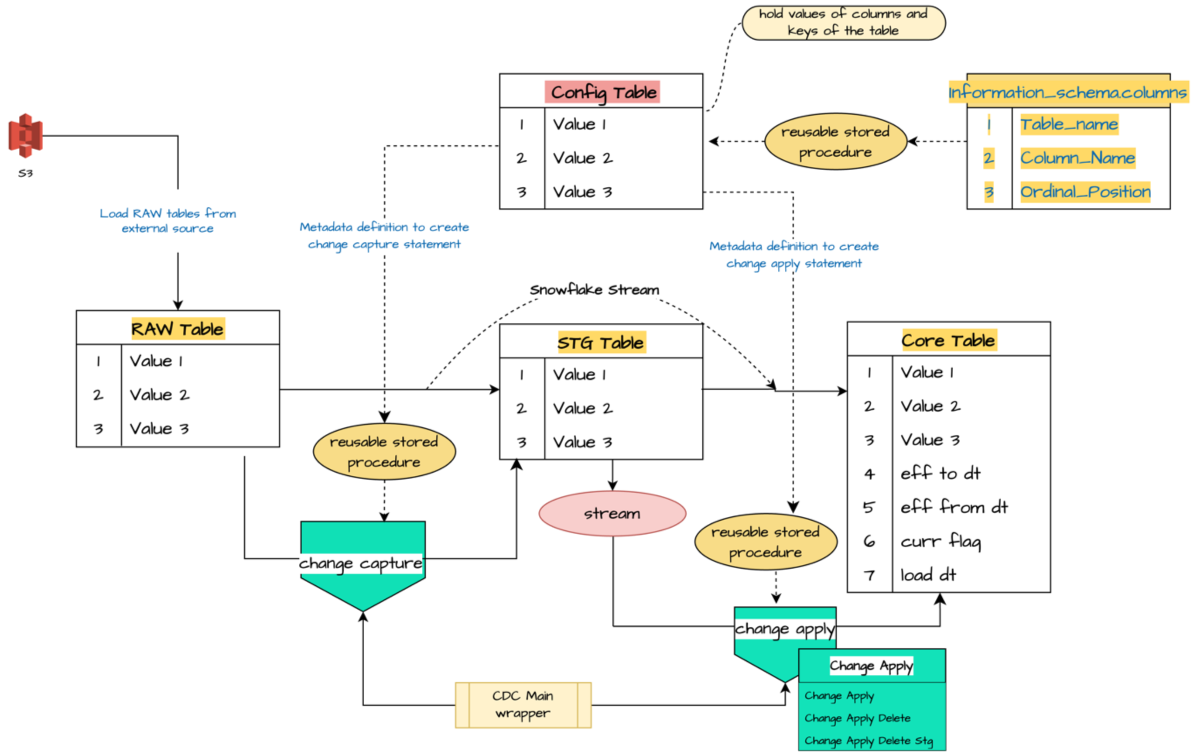
Snowflake Stream is an object in helping to identify the changes and mark the records as identifiable as new inserts/updates/deletes. The stream is to be created on top of the STG table and it will identify the changes with the help of SP_CHANGE_CAPTURE stored procedure.

The following diagram shows the end-to-end process flow of the data movement from the RAW table → STG Table → Core Table with respect to capturing the changes and applying the changes as part of Type 2 CDC process.

***Note**: The above codes/stored procedures will be attached with in the submission.*

© 2023 IBM Corporation

Reusable CDC Process to load data in Core Tables



3.0 Type 2 SCD Process

1. Create a Config table as follows. This config table will hold the details of the fields and Keys of the RAW and Core Tables and operations details like insert and updates. This table can be loaded through a stored procedure taking all the metadata information from the information schema. Use attached TableScripts.sql file to create the config table.

2. Create RAW/STG/TRFN by using attached TableScripts.sql file. Tables can be created for FULL/DELTA as per the demo requirements.

3. Create a stream on the STG table.

4. Create all the stored procedures mentioned above.

5. Execute the CRT_CDC_CONFIG stored procedure as shown in the following execution command. This stored procedure should be executed twice – one for RAW table and one for Core/Trfn table. After execution check the CDC config table records.

   - **Execution for RAW Table**

   CALL CRT_CDC_CONFIG ('<Database_Name>', '<Schema_Name>', '<RAW_Table_Name>', '<Key_Name>', '<Key_Type>', '<Table_Type>');

   1. ***Key_Type Value – 'MRG_STG_KY' in case processing configuration values of RAW Tables.***

   2. ***Key_Name – if more than one key then provide the keys separated with semi-colon (key1;key2).***

   3. ***Table_Type Value – 'R' in case of RAW table.***

   - **Execution for Core/Trfn Table**

CALL CRT_CDC_CONFIG ('<Database_Name>', '<Schema_Name>', '<RAW_Table_Name>', '<Key_Name>', '<Key_Type>', '<Table_Type>');

1. ***Key_Type Value – 'MRG_KY' in case processing configuration values of Core/Trfn Tables.***

2. ***Key_Name – if more than one key then provide the keys separated with semi-colon (key1;key2).***

3. ***Table_Type Value – 'T' in case of RAW table.***

6. Load the data in the RAW table with copy command or any customized stored procedure. In case of csv/delimited files directly files can be loaded in the RAW tables. In case of JSON/XML first the file needs to be flattened and then load into the RAW tables.

7. Run the following CDC Command to execute the wrapper script ()SP_CDC_MAIN) with parameters as mentioned with in the command line. The wrapper CDC Script (SP_CDC_MAIN) will sequentially execute SP_CHANGE_CAPTURE, SP_CHANGE_APPLY, SP_CHANGE_APPLY_DELETE and SP_CHANGE_APPLY_DELETE_STG.

   When source file processing is delta the SP_CDC_MAIN will execute SP_CHANGE_CAPTURE and SP_CHANGE_APPLY. The execution command is as follows:

   CALL SP_CDC_MAIN ('<RAW_Table_Name>','<STG_Table_Name>','<TRFN_Table_Name>','<Stream_Name>','<Database_Name>','<Source_Schema_Name>', '<Target_Schema_Name>', '<Load_Type>');

   ****LoadType – 'D' for Delta; 'F' for Full.***

   When source file processing is full the SP_CDC_MAIN will execute SP_CHANGE_CAPTURE and SP_CHANGE_APPLY and SP_CHANGE_APPLY_DELETE. The execution command is as follows:

   CALL SP_CDC_MAIN ('<RAW_Table_Name>','<STG_Table_Name>','<TRFN_Table_Name>','<Stream_Name>','<Database_Name>','<Source_Schema_Name>', '<Target_Schema_Name>', '<Load_Type>');

   ****LoadType – 'D' for Delta; 'F' for Full.***