## 

**Using MD5 function in Informatica for CDC technique**

## Purpose

Fortune 500 leading Telecom Company in US has extensive BI practice and uses Informatica tool for data extraction. MD5 (Message Digest Function) is a hash function in Informatica which is used to evaluate data integrity. The MD5 function uses Message-Digest Algorithm 5 (MD5) and calculates the checksum of the input value.MD5 is a one-way cryptographic hash function with a 128-bit hash value. MD5 returns a 32 character strings of hexadecimal digits 0-9 & a-f and returns NULL if the input is a null value.

## Challenge

If we have non-string columns (SMALLINT, INT, NUMBER, etc.) we need to convert them into characters using TO\_CHAR function because the MD5 Function only validates character strings.  If you don't use the TO\_CHAR, the output port using MD5 may become invalid.  Sometimes it does validate the output port using MD5, but the warning message generated may read validated and non-string data type (e.g. SMALLINT) columns are converted to string.  This warning has a great performance impact.

## Solution

Only update the changed records (any column change) in the target.  Instead of passing all existing records to the target for update, (whether changed or unchanged) it's always recommended to compare the records.

Column-to-column comparison is little painful if your column counts are high.  Concatenate all columns and use the MD5 Function (source and target records for the same key) and then compare the output from the MD5 Function.  The changed records can be identified and only those records can be updated into the target.  This calculation is done really fast and the output of MD5 Function can be used as a unique key to differentiate records

## Flow

To identify records for updates and inserts, we use a lookup transformation.  The cache built by lookup depends on two factors:

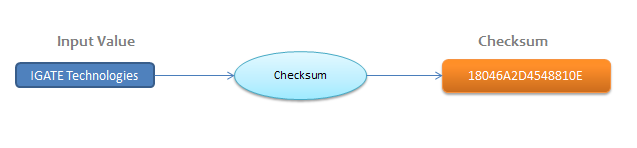
1. The number of columns in the comparison condition
2. The amount of data in the lookup table.

When there is not a primary key column to identify the changes, there are two options:

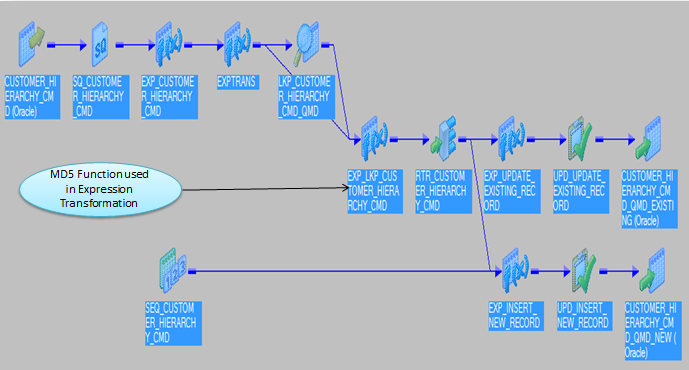
1. Compare all the columns in the lookup or...
2. Compare the data using the concept of power exchange change data capture.

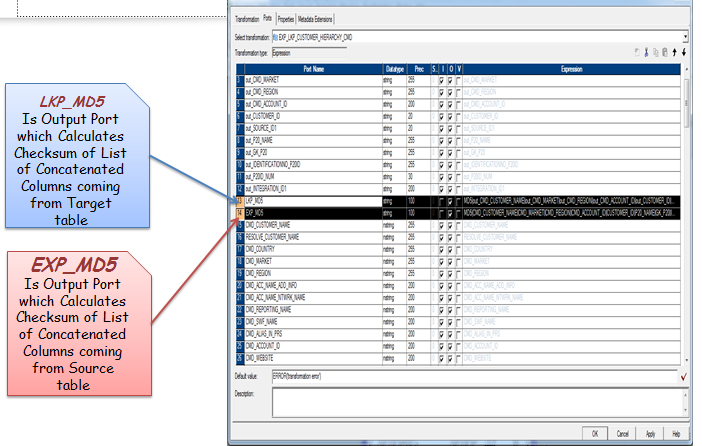
Beware that this can degrade the performance.  In this scenario, using the MD5 Function is our best option.

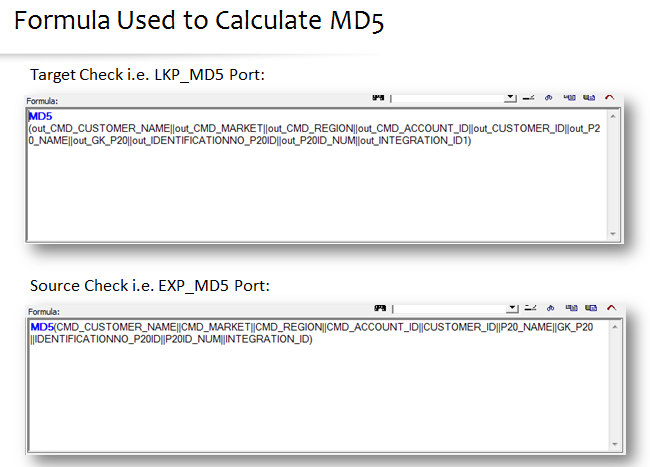
**Basic of Checksum**

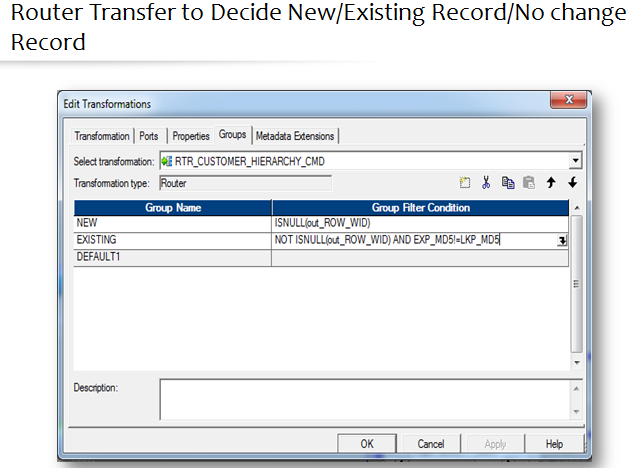


Basic Dataflow in Informatica:

MD5 function used in Expression transformation as Output Port







**Advantages**

1. MD5 will help in improving performance when compared to lookups only if the comparison columns are more than 10.  MD5 function enhances the performance as compared to lookups only when the comparison columns are more than 10.
2. Only update the changed records
3. Returns 32bit character string of hexadecimal digits 0-9 & a-f
4. Returns NULL if the input is null value
5. Reduce overall ETL run-time and cache memory usage, by caching only the required fields which are of utmost importance