使用iUAP DI进行基于快照的变更数据捕捉（CDC）

如果源表没有时间戳，没不能用触发器，就要用快照表了，通过比对获得数据的变化。快照表就是一次性抽取系统中全部数据，把这些数据加载到数据仓库的缓冲区中。下一次需要同步时，再从源系统中抽取全部数据，并把全部数据也放到数据仓库的缓冲区中，作为这个表的第二个版本，然后再比较这两个版本的数据，找到变化。例如，两列的表，列名是ID和Color。这个表的两个版本如下图：

快照1

|  |  |
| --- | --- |
| ID | COLOR |
| 1 | Black |
| 2 | Green |
| 3 | Red |
| 4 | Blu |

快照2

|  |  |
| --- | --- |
| ID | COLOR |
| 1 | Grey |
| 2 | Green |
| 4 | Blu |
| 5 | Yellow |

有很多方法可以获取这两个版本数据的差异。第一个办法就是在主键ID上做外连接并根据字段的比较结果增加一个标志字段（I表示插入，U表示更新，D表示删除，N表示没有变化），另外过滤掉没有变化的记录。

SELECT \* FROM

(SELECT CASE WHEN T2.ID IS NULL THEN 'D'

WHEN T1.ID IS NULL THEN 'I'

WHEN T1.COLOR <> T2.COLOR THEN 'U'

ELSE 'N'

END AS FLAG

, CASE WHEN T2.ID IS NULL THEN T1.ID

ELSE T2.ID

END AS ID

, T2.COLOR

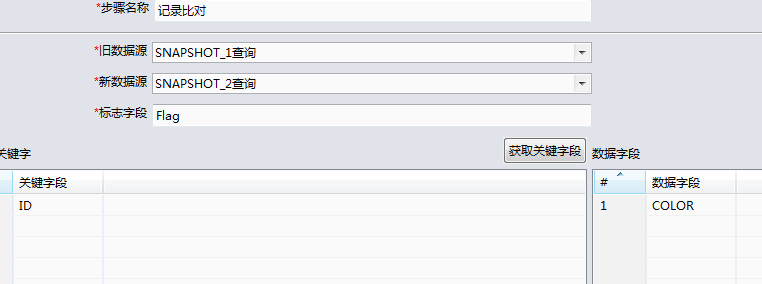
FROM SNAPSHOT\_1 T1 FULL OUTER JOIN SNAPSHOT\_2 T2 ON T1.ID = T2.ID) A

WHERE FLAG <> 'N'

结果如下表：

|  |  |  |
| --- | --- | --- |
| Flag | ID | COLOR |
| U | 1 | Grey |
| D | 3 | NULL |
| I | 5 | Yellow |

在iUAP DI中也可以比较表之间的差异，并增加一个标志字段表示I，U，D的变化，使用记录比对，就可以做同样的工作。设置如下图：



对比后的预览结果如下：

|  |  |  |
| --- | --- | --- |
| ID | COLOR | Flag |
| 1 | Grey | changed |
| 2 | Green | identical |
| 3 | Red | deleted |
| 4 | Blu | identical |
| 5 | Yellow | new |

可以根据Flag的值，利用switch/case 对Flag进行分流，进而对目标进行记录的删除、更新和插入。完整ETL过程见附件。

在实际应用中，快照表一般不允许建立在源数据库端和数据仓库中，最好建立一个缓冲数据库保存快照数据。

基于快照的CDC可以检测插入、更新和删除的数据，这是其优点，缺点是需要大量的存储空间来保存快照数据，另外数据量比较大时，有严重的性能问题。使用数据库的比较相对ETL引擎的性能要高，有时可以尝试第一种方法。