SAP DataServices 获取 CDC 数据的方法

变化数据捕获（CDC）技术是企业建设数据仓库，数据整合 ETL 工作中的重点和难点。企业需要定时将源业务系统中的数据抽取转换到数据仓库中，这些定期转换任务明显不可能每次把业务系统的数据全复制一次，那么如何把业务系统的变化数据，包括增加、删除、修改的数据，抽取转换到数据仓库里呢。

下面是一些常用的技术。

1、 时间戳

在没有时间戳的表上，如果新加入时间戳字段（Timestamp），则如果某些程序对改表进行插入操作，而且 Insert 语句中没有写列名，则会出现语法错误。

例如：

原表格 Tab_A 定义如下：

```
Create table_A (
A_ID int NOT NULL,
A_Desc varchar(20) NULL)
```

加入时间戳后变成

```
Create Table_A（
A_ID int NOT NULL,
A_Desc varchar(20) NULL,
Stamp datetime default getdate());
```

在表定义发生变化之后，如果使用如下语句则出现错误：

```
Insert into table_a values(1,’sdf’)
```

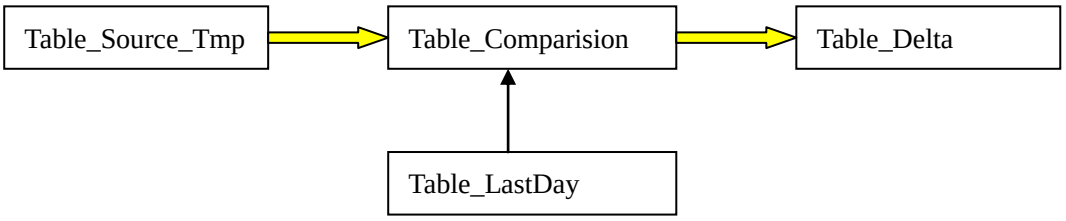
而此语句在表格更改之前是有效的。

因此，如果在生产系统的表格上增加时间戳字段，则需要修改应用程序，风险很高。故这种方法不建议采用。

2、 记录对比

假定生产系统中源表格名称为：Table_Source，在这个表格中的记录不断进行增、删、改的操作。

在判断增量之前，还需要把 Table_Source 表数据全部抽取到本地的 Table_Source_Tmp 中；并假定在前一天将该表中的全部数据都抽取到了本地，名为：Table_LastDay，则将 Table_Source_Tmp 和 Table_LastDay 进行对比，图示如下：



在 Table_Source_tmp 中如果从昨天到今天进行了 Insert 和 Update 操作，则 Table_Comparision 可以检测到，并输出到 Table_Delta 中。但如果 Delete 了记录，则 Delete 不能被检测到。

同事，在这个流程完成之后，还要清除 Table_LastDay 中的数据，然后把 Table_Source_Tmp 中的数据导入到 Table_LastDay 中，然后清除 Table_Source_Tmp 中的数据。

与使用时间戳方法类似，这种方法只能判断 Insert 和 Update 的变化，如果需要检测 Delete 操作，则需要以 Table_LastDay 作为 Input，Table_Source_Tmp 作为 Ref 重新比较，然后合并结果。

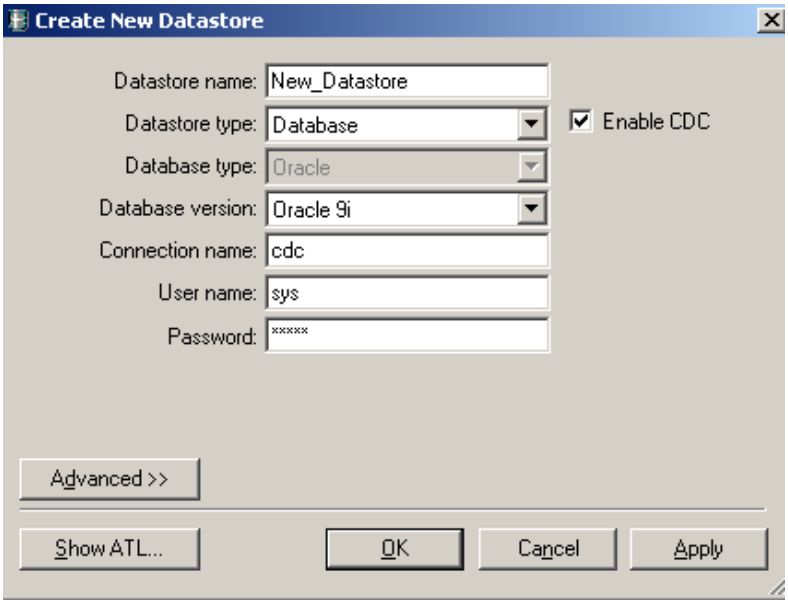
这种方法不适合大数据量的表格抽取，但可以用于代码表的增量判断。

3、 使用 DS 配合 Oracle 数据库自身 CDC 功能

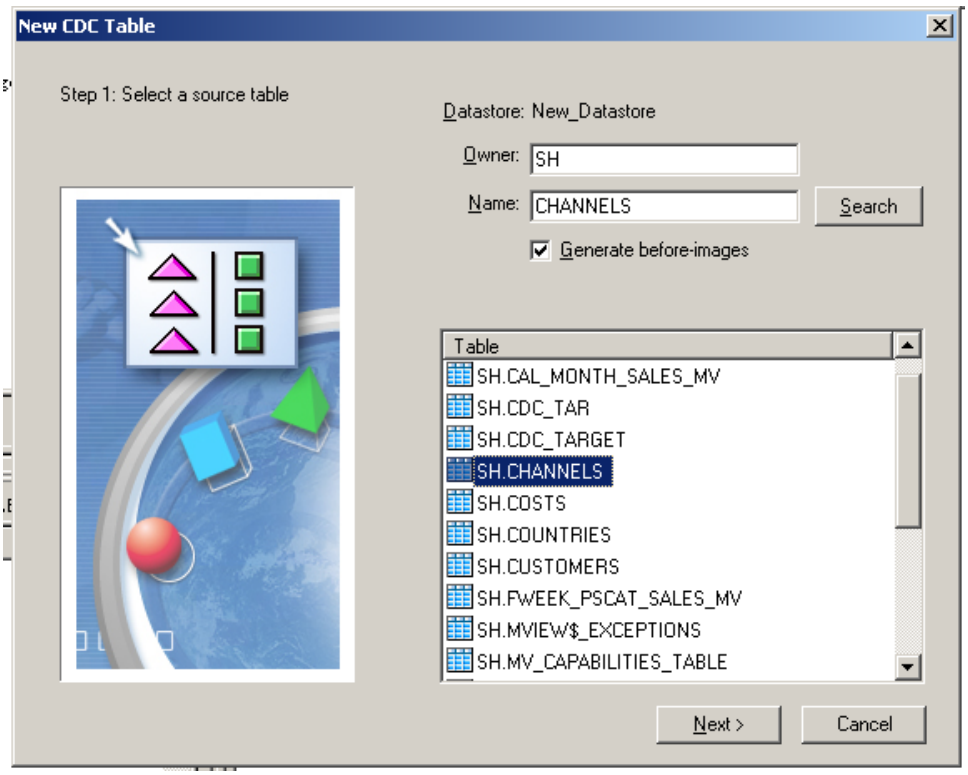
Oracle9i 和 Oracle10g 自身支持 CDC 功能。

在 DI 中如果需要使用 Oracle 本身的 CDC 功能，按照如下步骤进行操作：

1）创建 Oracle 数据库的 DataStore，在创建 DataStore 时，选择 CDC Enable 选项

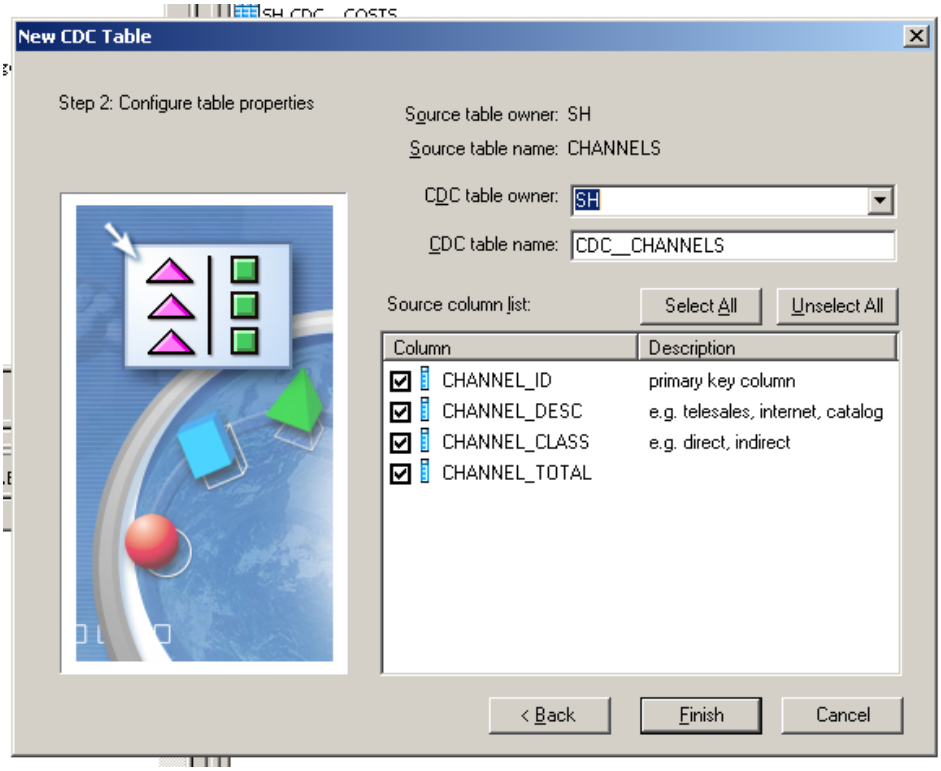


2）创建 Datastore 之后，如果 Oracle 数据库中已经存在 CDC 表，则可以直接 Import，如果没有 CDC 表格，则出现创建 CDC 表格的向导，如下图：



在 Owner 中填入 Schema 名称，点击 Search 按钮，则出现该 Schema 下的表格清单，选择要进行增量抽取的表格，例如 SH.CHANNELS 表。

3）点击 Next 按钮，选择要监控的列，只有被选择的列的数据发生变化时才会进入 CDC 表（此时 CDC 表还未创建），输入将要被创建的 CDC 表格的 Owner 名称和表名称，如下图：



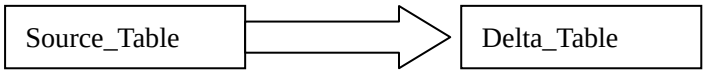
4）点击 Finish 按钮，CDC 表格被创建起来并被导入到该 Datastore 中。

5）在数据库中，对 SH.CHANNELS 进行的操作（Insert/Update/Delete）将被记录到新创建

的 CDC 表中，这里是被记录到 CDC__CHANNELS 中。
6）如果使用 CDC__CHANNELS 作为数据源实现增量抽取，则需要另外创建一个 DataStore，不选择 CDC Enable 选项，并且把 CDC__CHANNELS 当作普通表格导入即可。
注意在每次抽取 CDC__CHANNELS 数据之后及时把数据从表中删除。

4、 在源系统上创建 Trigger

在源表上创建 Trigger，可以捕获员表上的任何操作变化，将其输入到另外一个表格中（称为增量表），抽取程序取走增量表的数据后，马上清空增量表。



理论上，建立 Trigger 方法能够实现与 CDC 表一样的功能，只是建立 Trigger 需要在数据库上写脚本，方法不如用 DI 创建 CDC 表格方便。