A Simple ETL Process for Business Calendar

Yifu Huang

Recently I have worked on the Business Day Calendar ETL and have learned some knowledge on a simple ETL process. I think it is highly reusable especially loading part and want to share it with you.

If you need build table from scratch and load data into table especially large volume data, try to have a look at the following simple ETL process.

1. **Define the table structure**

   The first thing is writing data definition language to define the table structure what I want to extract. In a word, I need write CREATE TABLE clause to clarify all column types.

2. **Extract and transform data**

   Because I build the Business Day Calendar from scratch, I need to extract data from HTML and then transform it. The general method is writing regular expressions to extract data and then writing string functions to transform data in some program language.

3. **Load data**

   Teradata provides some powerful utilities such as BTEQ IMPORT, FASTLOAD and MULTILOAD to help user load data. BTEQ IMPORT and FASTLOAD samples below.

   3.1 BTEQ IMPORT script:

```
.LOGON mozart/username,password;
CREATE TABLE P_CSI_TBS_T.BTEQ_IMPORT_TEST(
   CNTRY_ID DECIMAL(4,0) NOT NULL, -- joins to DW_COUNTRIES
   CAL_DT DATE FORMAT 'yyyy/mm/dd' NOT NULL, -- joins to DW_CAL_DT
   HOLIDAY_DESCR VARCHAR(64) NOT NULL, -- in English
   SOURCE_URL VARCHAR(256), -- only offical government sources
   LEAPYR_DT_IND CHAR(1) NOT NULL DEFAULT 'N',
   -- 'Y', if the date is the 29th February in a leapyear
   LST_MDF_USER CHAR(10) CHARACTER SET LATIN NOT CASESPECIFIC NOT NULL,
   -- enter name of creator here
   LST_MDF_DATE DATE FORMAT 'yyyy/mm/dd' NOT NULL DEFAULT CURRENT_DATE,
   -- date when record is created or changed
   CAL_DT_FORMER DATE FORMAT 'yyyy/mm/dd'
) PRIMARY INDEX (CNTRY_ID, CAL_DT);
.IMPORT VARTEXT ' ' FILE=C:\Users\username\Desktop\DATA.txt;
.REPEAT *
USING
CNTRY_ID (VARCHAR(4)),
CAL_DT (VARCHAR(10)),
HOLIDAY_DESCR (VARCHAR(64)),
SOURCE_URL (VARCHAR(256)),
LEAPYR_DT_IND (VARCHAR(1)),
LST_MDF_USER (VARCHAR(10)),
LST_MDF_DATE (VARCHAR(10)),
CAL_DT_FORMER (VARCHAR(10))
```

```
INSERT INTO P_CSI_TBS_T.BTEQ_IMPORT_TEST
(CNTRY_ID,CAL_DT,HOLIDAY_DESCR,SOURCE_URL,
LEAPYR_DT_IND,LST_MDF_USER,LST_MDF_DATE,CAL_DT_FORMER)
VALUES
(
CAST(:CNTRY_ID AS DECIMAL(4,0)),
CAST(:CAL_DT AS DATE FORMAT 'YYYY/MM/DD'),
:HOLIDAY_DESCR,
:SOURCE_URL,
:LEAPYR_DT_IND,
:LST_MDF_USER,
CAST(:LST_MDF_DATE AS DATE FORMAT 'YYYY/MM/DD'),
CAST(:CAL_DT_FORMER AS DATE FORMAT 'YYYY/MM/DD')
);
.LOGOFF
.QUIT
```

3.2 BTEQ IMPORT usage:

To execute this script, first save this script as **BTEQ_IMPORT_SCRIPT.txt**, then type this command **BTEQ < C:\Users\username\Desktop\BTEQ_IMPORT_SCRIPT.txt > C:\Users\username\Desktop\BTEQ_IMPORT_LOG.txt** in cmd.exe, and finally execute it. (Path in command should be changed if you want to try)

3.3 BTEQ IMPORT note:

3.3.1 Custom delimiter can be define after the key word .IMPORT VARTEXT. (While using tab, try ‘ ’ rather than ‘\t’)

3.3.2 The key word VARTEXT means all Using definition should be VARCHAR, VARBYTE.

3.3.3 BTEQ IMPORT is suitable for small data volume, because it inserts records one by one. So is Teradata SQL Assistant Import.

3.4 FASTLOAD script:

```
logon mozart/username,password;
CREATE TABLE P_CSI_TBS_T.FASTLOAD_TEST(
   CNTRY_ID DECIMAL(4,0) NOT NULL, -- joins to DW_COUNTRIES
   CAL_DT DATE FORMAT 'yyyy/mm/dd' NOT NULL, -- joins to DW_CAL_DT
   HOLIDAY_DESCR VARCHAR(64) NOT NULL, -- in English
   SOURCE_URL VARCHAR(256), -- only offical government sources
   LEAPYR_DT_IND CHAR(1) NOT NULL DEFAULT 'N',
   -- 'Y', if the date is the 29th February in a leapyear
   LST_MDF_USER CHAR(10) CHARACTER SET LATIN NOT CASESPECIFIC NOT NULL,
   -- enter name of creator here
   LST_MDF_DATE DATE FORMAT 'yyyy/mm/dd' NOT NULL DEFAULT CURRENT_DATE,
   -- date when record is created or changed
   CAL_DT_FORMER DATE FORMAT 'yyyy/mm/dd'
) PRIMARY INDEX (CNTRY_ID, CAL_DT);
set record vartext "        ";
define
```

```
CNTRY_ID (VARCHAR(4)),
CAL_DT (VARCHAR(10)),
HOLIDAY_DESCR (VARCHAR(64)),
SOURCE_URL (VARCHAR(256)),
LEAPYR_DT_IND (VARCHAR(1)),
LST_MDF_USER (VARCHAR(10)),
LST_MDF_DATE (VARCHAR(10)),
CAL_DT_FORMER (VARCHAR(10))
file=C:\Users\username\Desktop\DATA.txt;
show;
begin loading P_CSI_TBS_T.FASTLOAD_TEST
errorfiles P_CSI_TBS_T.error_test_1, P_CSI_TBS_T.error_test_2;
INSERT INTO P_CSI_TBS_T.FASTLOAD_TEST
(CNTRY_ID,CAL_DT,HOLIDAY_DESCR,SOURCE_URL,
LEAPYR_DT_IND,LST_MDF_USER,LST_MDF_DATE,CAL_DT_FORMER)
VALUES
(
:CNTRY_ID,
:CAL_DT,
:HOLIDAY_DESCR,
:SOURCE_URL,
:LEAPYR_DT_IND,
:LST_MDF_USER,
:LST_MDF_DATE,
:CAL_DT_FORMER
);
end loading;
logoff;
```

3.5 FASTLOAD usage:

To execute this script, first save this script as **FASTLOAD_SCRIPT.txt**, then type this command **FASTLOAD < C:\Users\username\Desktop\FASTLOAD _SCRIPT.txt > C:\Users\username\Desktop\FASTLOAD_LOG.txt** in cmd.exe, and finally execute it. (Path in command should be changed if you want to try)

3.6 FASTLOAD note:

3.6.1 Custom delimiter can be define after the key word set record vartext. (While using tab, try “ “ rather than “\t”)

3.6.2 The key word VARTEXT means all Using definition should be VARCHAR, VARBYTE.

3.6.3 The name of errorfiles should not exist in database already.

3.6.4 FASTLOAD is suitable for large data volume.

Sample data is also available if you want you can have a try.

Besides, as to the Business Day Calendar, We’ve finalized it for US/UK/DE/CN/HK. The data comprises holidays from 2000 to 2015 (government resources wherever possible) and includes special weekend working days for CN.