無料で多機能な OSS の ETL ツール「Kettle」を使ってみよう！(2)

情報政策課　技術専門職員　金森　浩治

１．はじめに

富山大学総合情報基盤センター広報 vol.12 にて，OSS の ETL「Kettle」の機能と基本的な使用方法を紹介いたしました。

今回は、「値マッピング」や「分岐」、「計算」といった機能についてご紹介いたします。

2. 用語説明

2.1　OSS とは？

OSS とは Open Source Software の略で、ソースコードが公開されているソフトウェアのことです。

OSS 製品は無料で使用できるものが多いのが特徴です。

OSS で有名なものとして、Web ソフトウェア”Apache”、アプリケーションサーバソフトウェア”Tomcat”などがあります。

2.2 ETL ツールとは？

「ETL」とは、データベースや Web サービスなどのデータソースからデータを取得し、適切な形にデータ変換し、データベース等にデータを挿入するツールです。

なお「ETL」は Extract/Transform/Load の頭文字をとった略称です。各々の単語の意味は次の通りです。

Extract ・・・・データ抽出
Transform・・・変換
Load ・・・・・データ挿入

Extract(データ抽出)、Transform(変換)、Load(データ挿入)の詳しい説明については、富山大学総合情報基盤センター広報 vol.12「無料で多機能な OSS の ETL ツール『Kettle』を使ってみよう！」をご参照ください。

2.3 Kettle とは？

Kettle は BI スイーツ”Pentaho”の一部です。CE 版は OSS で提供されており、無料で使用できます。

Kettle のインストール方法や基本的な使用法については、富山大学総合情報基盤センター広報 vol.12 「無料で多機能な OSS の ETL ツール『Kettle』を使ってみよう！」をご参照ください。

なお本稿ではバージョン 6.0.1.0-386 を使用します。そのため最新バージョンとは画面等の差異がありますので、ご了承ください。

3. 機能紹介

今回は「値マッピング」や「分岐」、「計算」といった機能についてご紹介いたしますが、共通のデータソースとして、以下データが記載された Excel ファイルを読み込み、処理しております。

学生番号	姓	名	性別
1001	富山	太郎	M
1002	高岡	佐紀	F
1003	魚津	次郎	M
2001	射水	かおり	F

表 1　学生マスタ

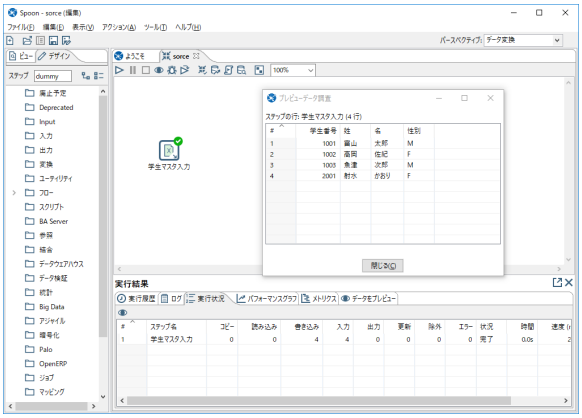

図 1　学生マスタ読込み結果

3.1 値マッピング

“値マッピング”機能とは、特定のデータを指定したデータに変換する処理する機能です。

例えば、「表 1 学生マスタ」の場合だと、性別を「男」や「女」に変換する場合に使用します。

設定方法ですが、「値マッピング」機能は”変換”フォルダにありますので、画面右にドラッグアンドドロップします。

続いてシフトキーを押しながら、”データソース”を左クリックしたまま、”値マッピング”上で左クリックを離し、”データソース”と”値マッピング”を図 2 のように連結させます。

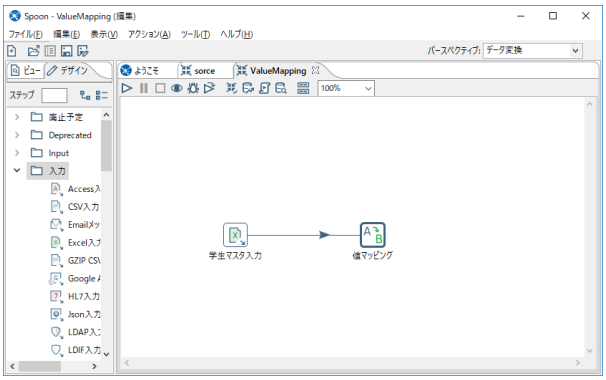

図 2 ”データソース”と”値マッピング”の連結

”値マッピング”をダブルクリックすると設定画面が表示されますので、図 3 のように設定してください。

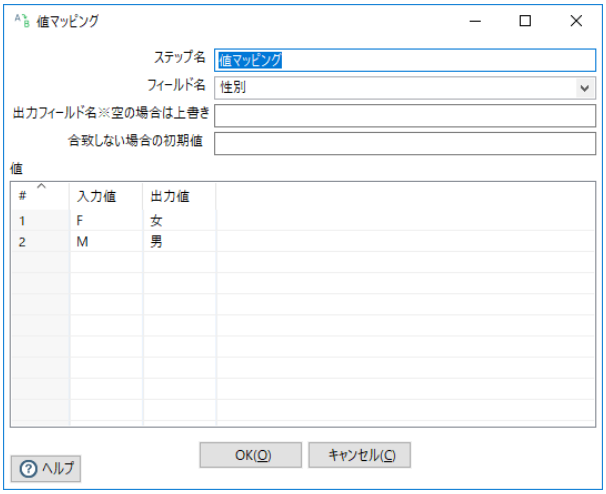

図 3 “値マッピング”の設定

では、実行結果を確認しましょう。確認方法は、”値マッピング”を右クリックし、メニューから「プレビュー」を選択します。続いて「クイック起動」ボタンをクリックします。すると図 4 のように”値マッピング”の実行結果が表示されます。

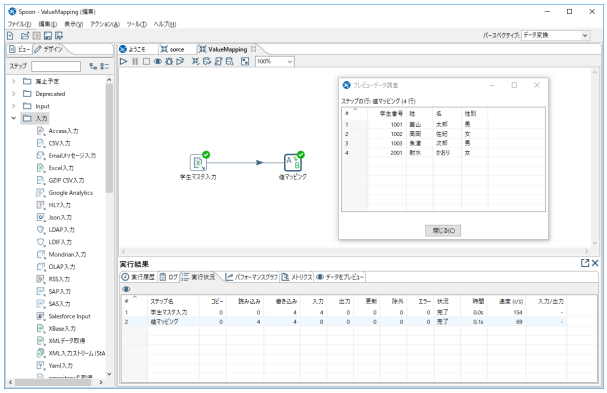

図 4 ”値マッピング”の実行結果

性別欄が漢字表記に変わっていることが確認できます。

3.2 分岐

“分岐”機能はプログラムで言うところの”If”や”Select case”にあたる処理です。値によって処理を変えたい場合に使用します。

まずは”If”に該当する”フィルター”機能から説明します。”フィルター”機能は「フロー」フォルダにありますので、”値マッピング”と同様に図 5 のように設定してください。

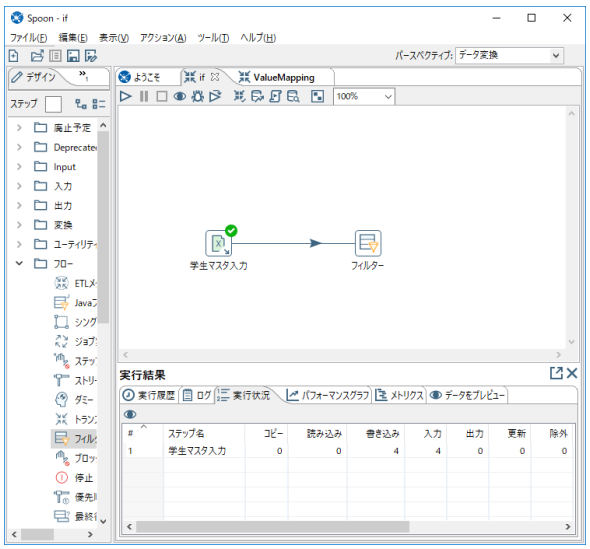

図 5 ”データソース”と”フィルター”の連結

続いて”フィルター”をダブルクリックして図6のように設定してください。

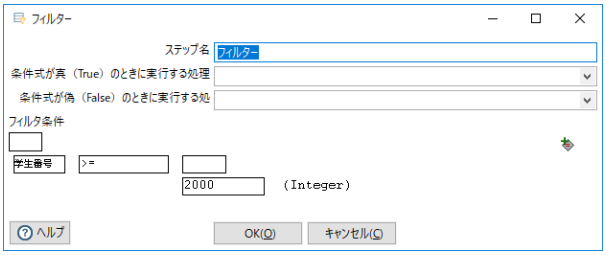

図 6 “フィルター”の設定

それでは実行結果を確認しましょう。”フィルター”を右クリックし、「プレビュー」メニューを選択し、「クイック起動」をクリックします。

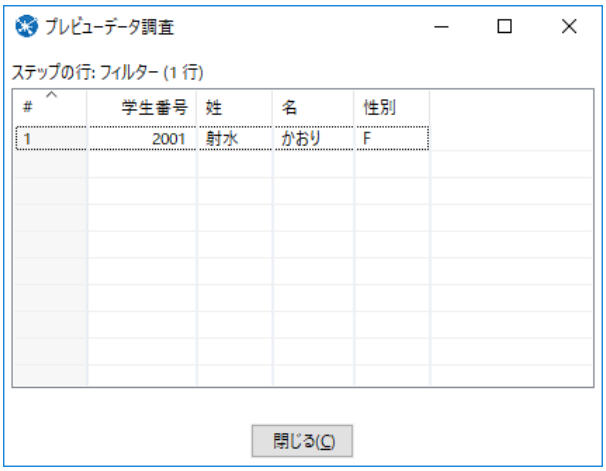

図 7 “フィルター”の実行結果(true 側)

学生番号が 2000 以上のデータが表示されていることがわかります。しかしこれだと 2000 未満のデータについては処理することができません。

そのため、フローフォルダにある”ダミー”を 2 つ配置し、フィルターから連結させてみてください。以下のように表示されますので片方には”true”、もう一方には”false”を設定してください。

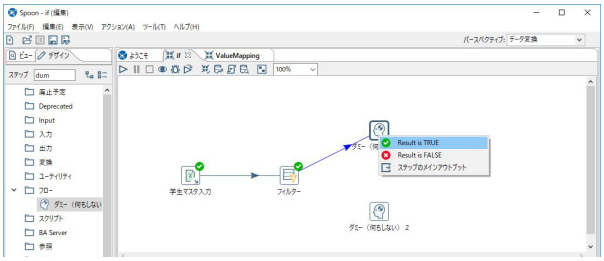

図 8 “フィルター”による分岐の設定

”false”側のダミーのプレビュー結果を見ると学生番号が 2000 未満のデータが表示されていることが確認できるかと思います。

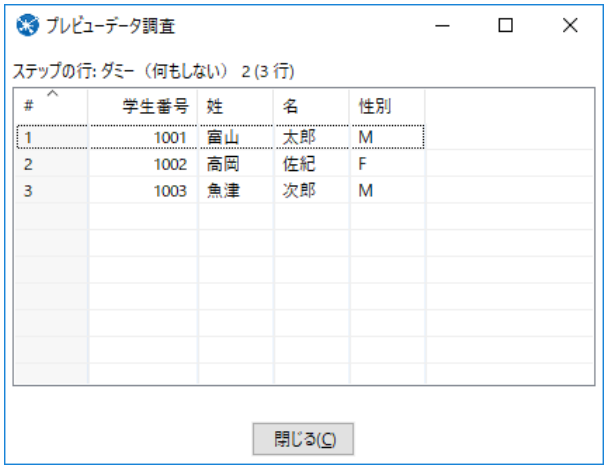

図 9 “フィルター”の実行結果(false 側)

このようにしてデータを設定した条件でデータを振り分け、処理をデータごとに変更することが可能です。

また分岐が複数にわたる場合、”条件分岐(Switch / Case)”が利用できます。

図 10 は”条件分岐(Switch / Case)”の設定画面で、学生番号ごとに分岐させる内容となっております。

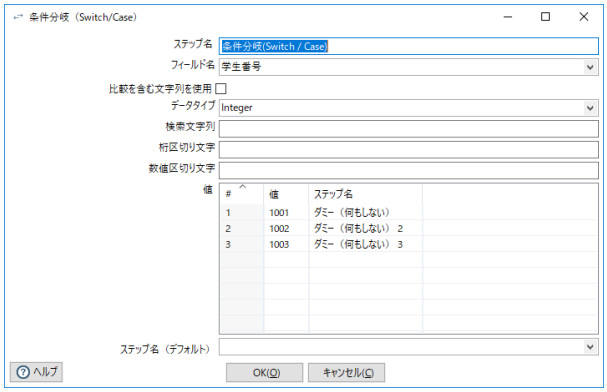

図 10 条件分岐(Switch / Case)”の設定画面

3. 3 計算

ここでは“計算”機能を用いた文字列連結について説明します。”計算”は「変換」フォルダにありますので、図 11 のように設定してください。

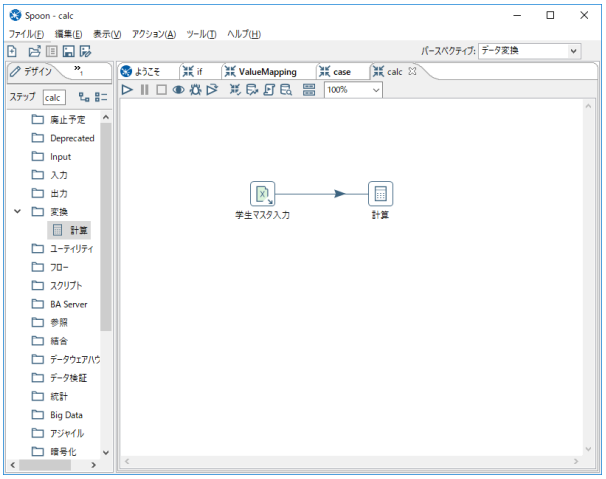

図 11 ”データソース”と”計算”の連結

続いて” 計算” をダブルクリックし、図 12 のように設定しましょう。

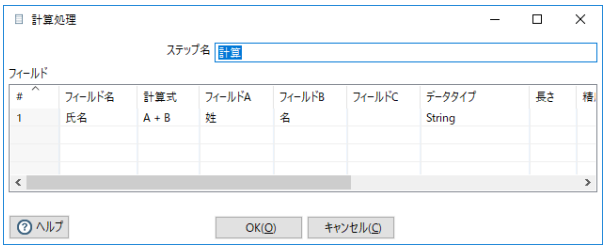

図 12 ”計算”の設定画面

”計算”のプレビュー結果を確認すると、図 13 のように、姓と名が連結されていることがわかります。

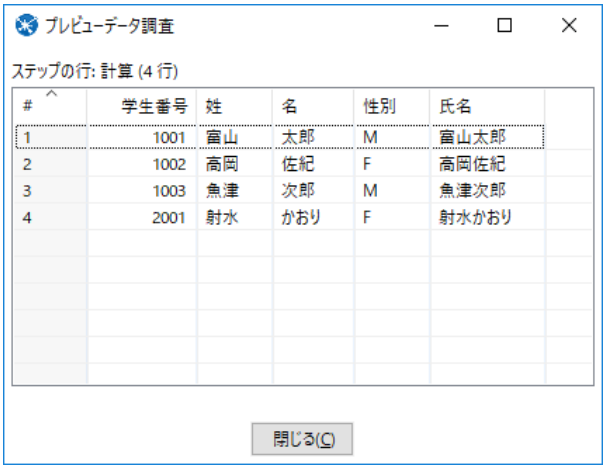

図 13 ”計算”の実行結果

“計算” で処理できる内容は文字列の連結以外にもたくさんあります。図 14 は “計算” でできる処理の一例です。

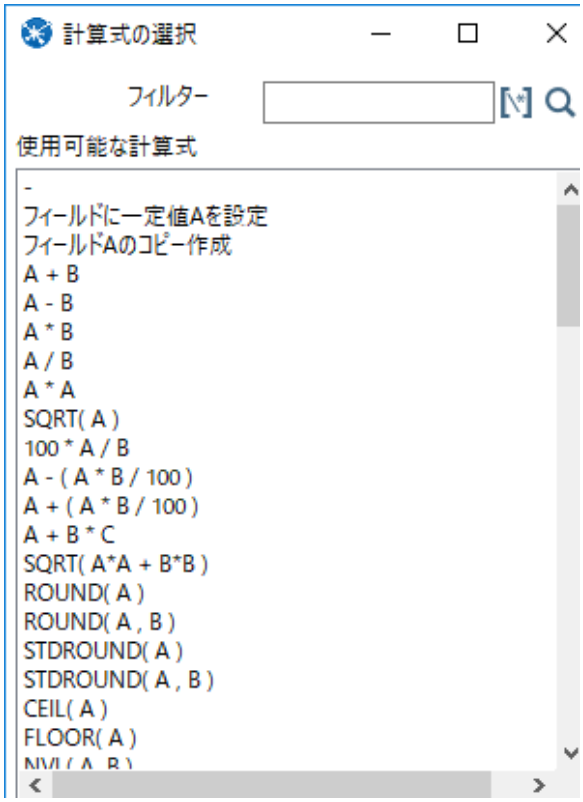

図 14 “計算式”の選択画面

4. 最後に

以上で簡単に説明を終えますが、本来はもっと複雑な変換をします。興味がある方は、data-integration¥samples フォルダ配下にサンプルファイルが多数ありますので、参考にしてください。

参考文献

1) 富山大学総合情報基盤センター広報 vol.12, P65-72.