無料で多機能な OSS の ETL ツール「Kettle」を使ってみよう！

情報政策課　技術職員　金森　浩治

１．はじめに

データ処理を行うにあたって非常に便利なツール”ETL”。本稿では OSS の ETL「Kettle」の機能とその使用方法を紹介します。

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

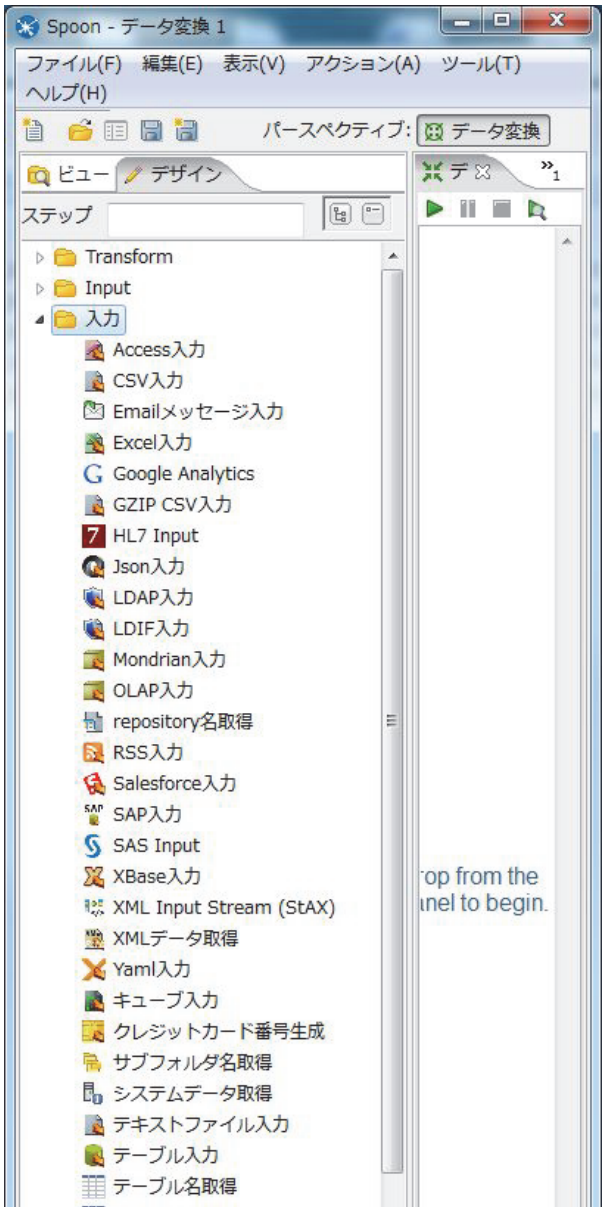

図 1 データ源の種類

2.2.1 Extract（データ抽出）

ファイルや DB ベース、Web サービスといった各種データ源からデータを取得する工程です。Kettle の場合、図 1 のようなデータ源を使用できます。

2.2.2 Transform(変換)

抽出したデータを目的の形に変換・加工する工程です。

図 2, 3 は Kettle で使用できる変換および参照の一例です。

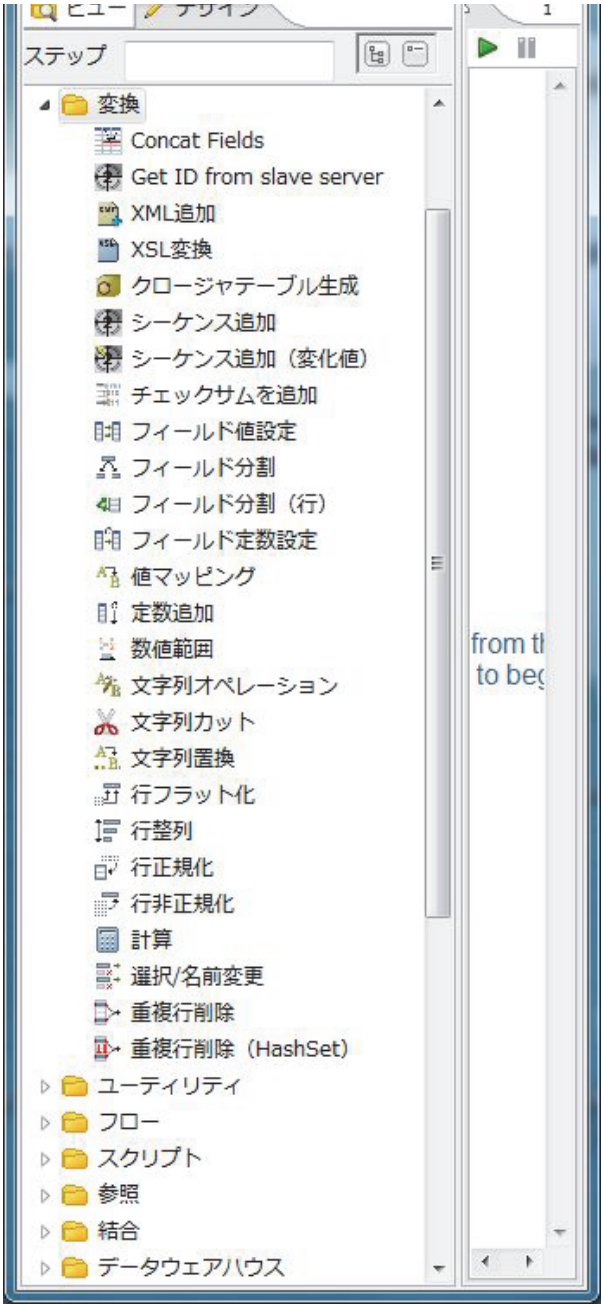

図 2 変換

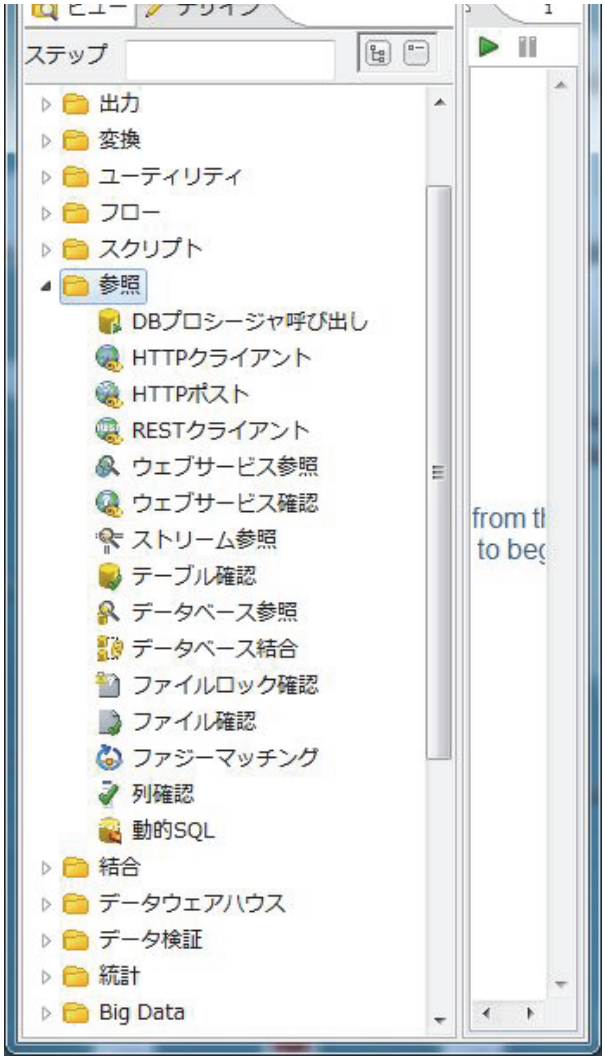

図 3 参照

2.2.3 Load(データ挿入)

データをデータベースや XML、LDAP 等に出力する工程です。Kettle の場合、図 4 のような形に出力できます。

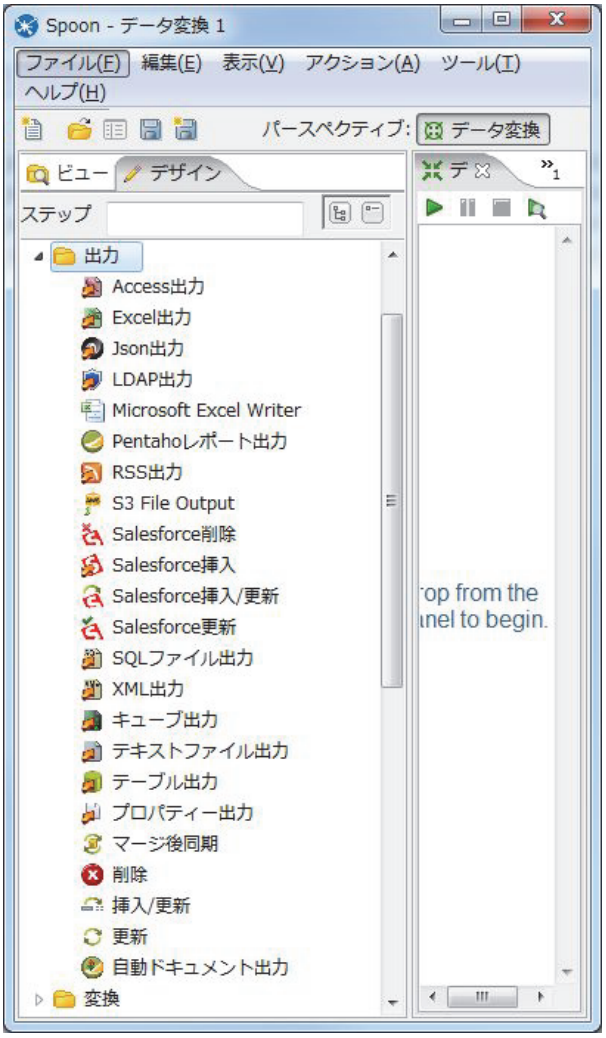

図 4　出力

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。

３．使ってみよう！

さっそく Kettle を使ってみましょう。本稿では以下のやり方を説明します。

・CSV データを Excel に変換する
・Excel ファイルを連結する

3.1 Kettle のインストール

最初に kettle をインストールする必要があります。手順は以下の通りです。

1. java のインストール
2. path の設定
3. Kettle のダウンロードし、解凍
4. 解凍フォルダを C ドライブ直下に保存

「java のインストール」や「path の設定」がわからない人は google 等で検索してみてください。

また Kettle のダウンロードサイトについても google 等で検索するとヒットすると思います。

3.2 CSV データを Excel に変換してみよう

試しに CSV ファイルを Excel に変換してみましょう。

通常であれば、CSV ファイルを Excel で開いて Excel で保存すればできますが、その場合、CSV データに改行が入ってたり、”00054”といった文字の場合、うまくいかないことがあります。こういった場合でも、Kettle を用いればうまく変換できます。

1. CSV ファイルを作る

以下のような内容が書かれている CSV ファイルを作ってデスクトップ等に保存してください。

a,011
b,022
c,033

2. C:¥data-integration¥Spoon.bat をダブルクリックします。

3. メニューより[ファイル]→[新規]→[データ変換]を実行 します。

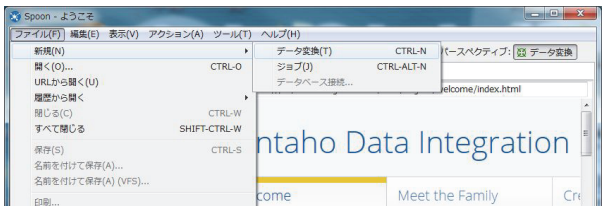

図 5

4. [入力]→[CSV 入力]を右エリヤにドラック&ドロップし、図 6 のようにします。

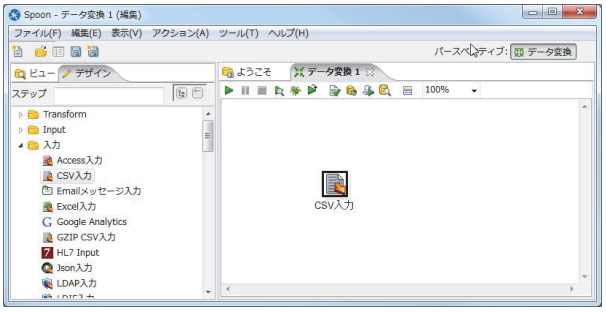

図 6

5. [出力]→[Excel 出力]を右エリヤにドラック&ドロップし、図 7 のようにします。

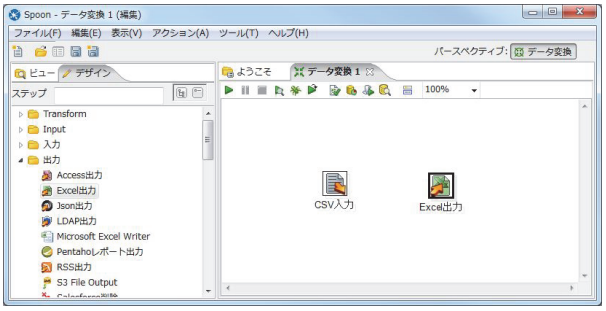

図 7

6. Shift キーを押しながら”CSV 入力”アイコン上で左クリックしながらを”Excel 出力”アイコン上で離すと図 8 のように矢印が作成されます。

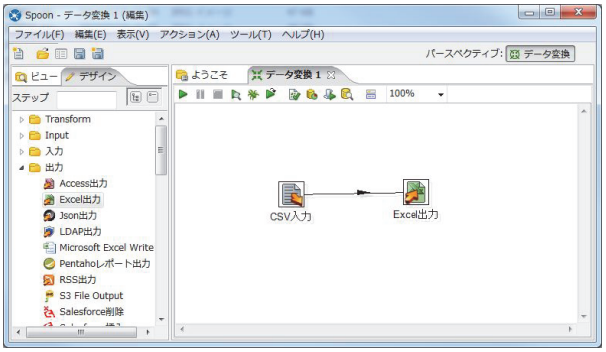

図 8

7. “CSV 入力”のアイコンをダブルクリックして、参照ボタンをクリックし、手順 1 で作成した CSV ファイルを選択してください。また、「ヘッダー・レコードを含む」チェックボックスのチェックを外し、下の表に[1][2]のように入力します。

入力後「OK」ボタンをクリックし画面を閉じます。

[1]
フィールド名：Field_000
データ・タイプ：String
[2]
フィールド名：Field_001
データ・タイプ：String

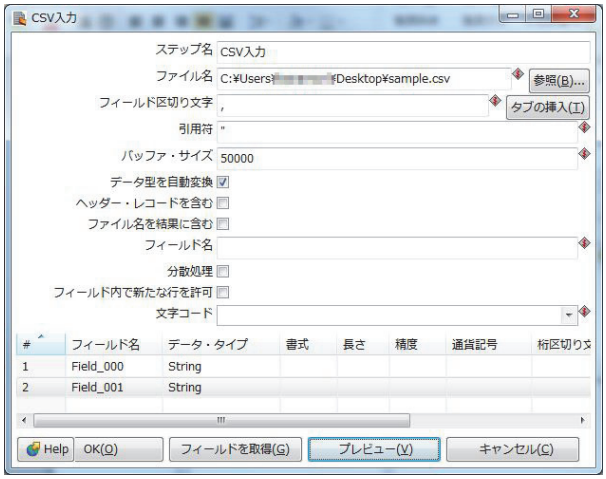

図 9

8. “Excel 出力”のアイコンをダブルクリックして、ファイル名欄に Excel ファイルの出力先を入力してください。

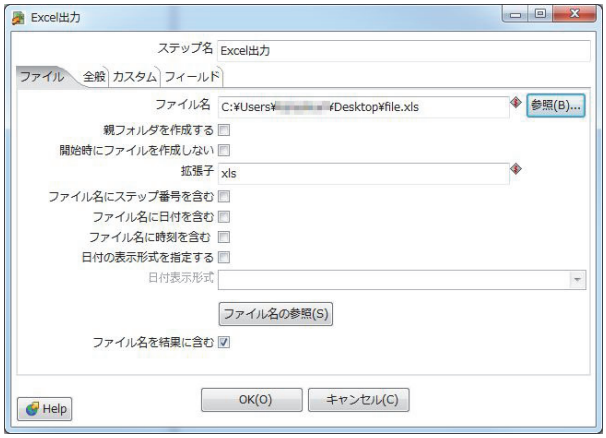

図 10

また、「フィールド」タブを選択し、「フィールドを取得」ボタンをクリックしてください。すると”フィールド名”と”データタイプ”が入力されます。そして「OK」ボタンをクリックして画面を閉じます。

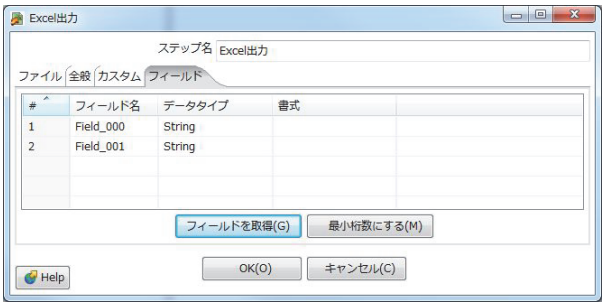

図 11

9. 実行ボタン(図 12 参照)をクリックしてください。

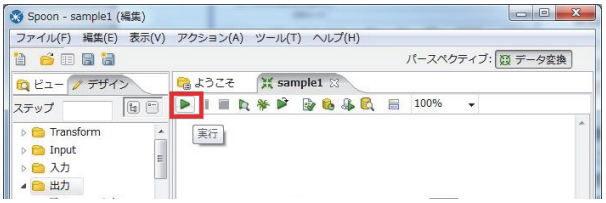

図 12

以下の画面が表示されるので「実行」ボタンをクリックしてください。

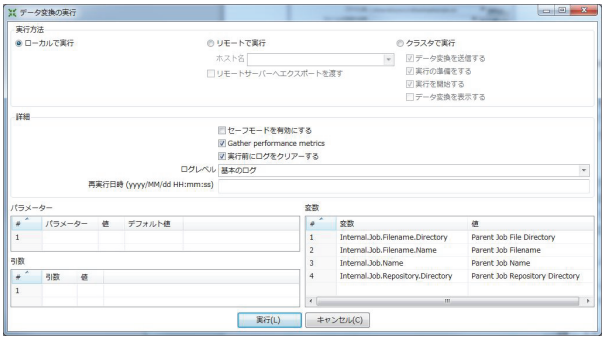

図 13

すると Excel 出力先に Excel ファイルが出来ています。開くと下記のようになっています。

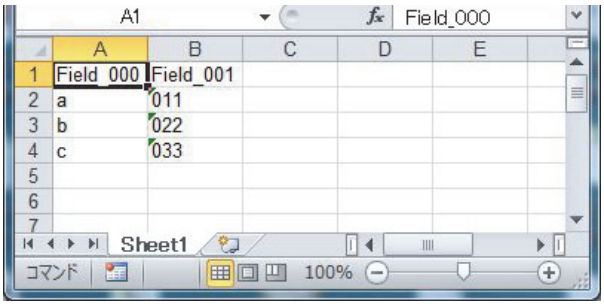

図 14

3. 3 Excel データを連結してみよう

図 15, 16 のような「学生ファイル」と「成績ファイル」を連結してみます。

[学生ファイル]
ファイル名：学生マスタ.xlsx

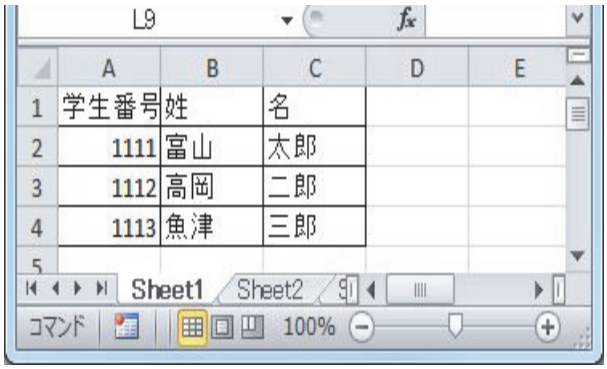

図 15

[成績ファイル]
ファイル名：成績トランザクション.xlsx

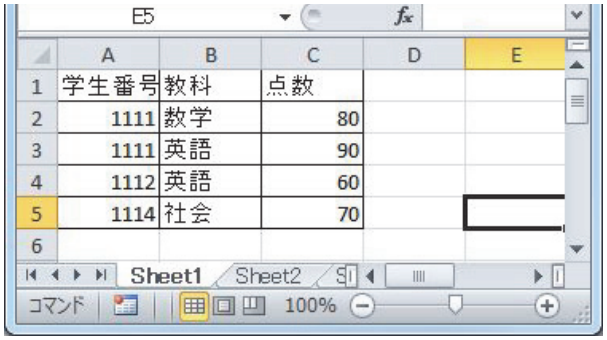

図 16

1. C:¥data-integration¥Spoon.bat をダブルクリックします。

2. メニューより[ファイル]→[新規]→[データ変換]を実行します。

3. 画面左より[入力]→[Excel 入力]を右エリヤにドラック&ドロップし、図 17 のようにします。

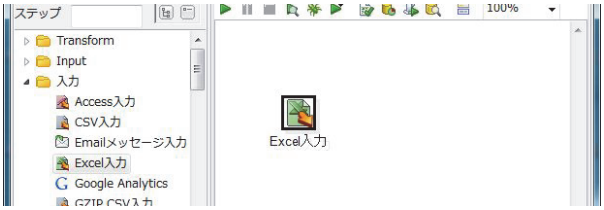

図 17

4. [Excel 入力]アイコンをダブルクリックし、「スプレッドシートタイプ(エンジン)」を”Excel 2007 XLSX (Apache POI)”を選択、「ファイル名のリスト」欄に、”学生マスタ.xlsx”ファイルのフルパスを入力してください。

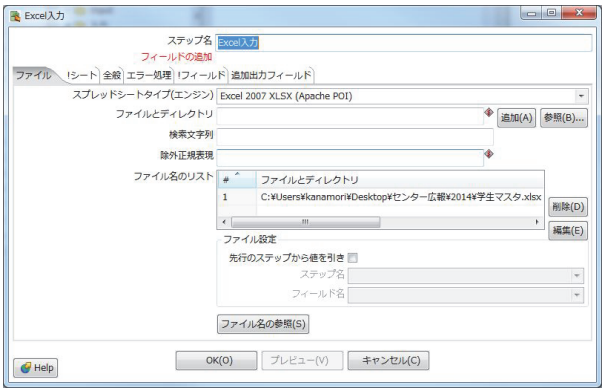

図 18

5. 「シート」タブをクリックし、「シート名」欄に”Sheet1”と入力してください。

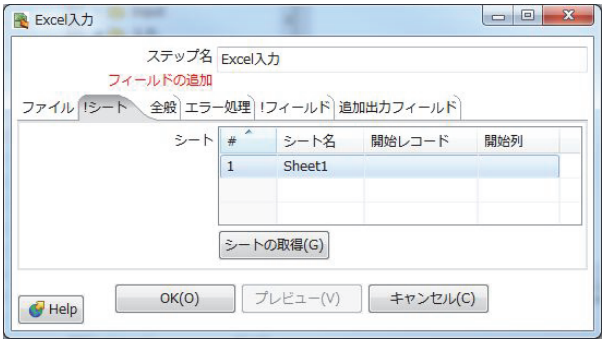

図 19

6. 「フィールド」タブをクリックし、「フィールドの取得」ボタンをクリックするとフィールド名等が入力されますので、「OK」ボタンをクリックしてください。

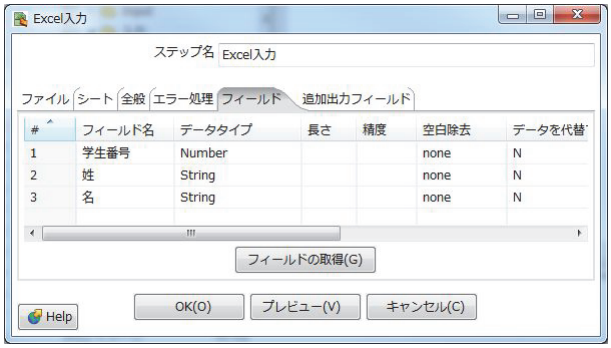

図 20

7. 同様に手順 3～6 を「成績ファイル」に対しても行ってください。(下図は「フィールド」タブの入力内容)

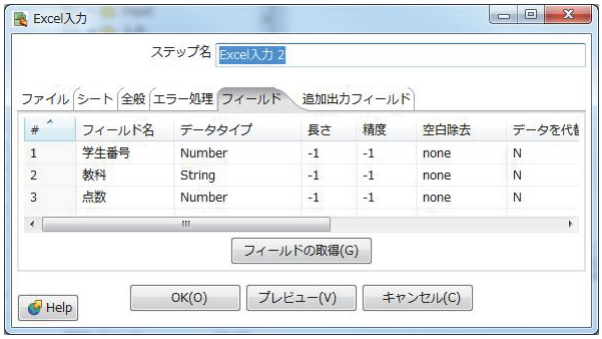

図 21

8. 画面左より[結合]→[マージ結合]を右エリヤにドラック&ドロップし、Shift キーを押しながら”Excel 入力”および”Excel 入力 2”アイコン上で左クリックしながら” マージ結合”アイコン上で離し、図 22 のような矢印を作成します。

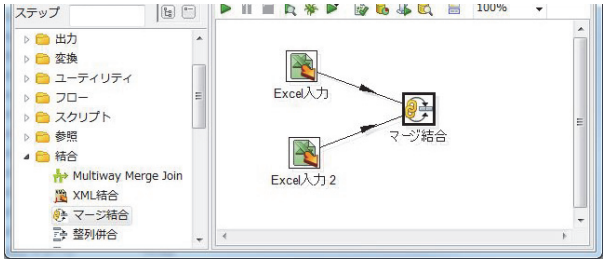

図 22

9. 「マージ結合」アイコンをクリックし、以下のように入力し「OK」ボタンをクリックしてください。

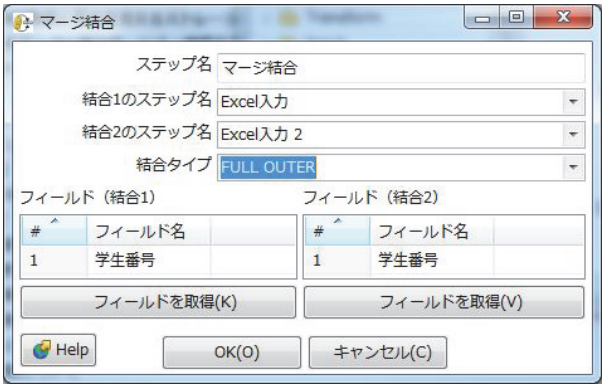

図 23

なお「OK」ボタンクリック後、警告のポップアップが表示されますが、了解ボタンをクリックしてください。

10. 画面左より[出力]→[Excel 出力]を右エリヤにドラック&ドロップし、Shift キーを押しながら”マージ結合”アイコン上で左クリックしながら”Excel 出力”アイコン上で離し、図 24 のような矢印を作成します。

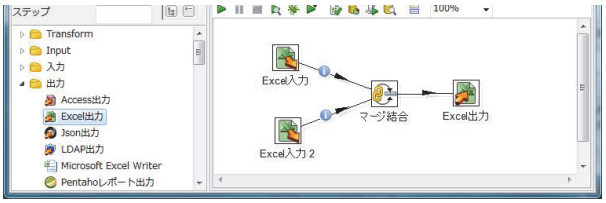

図 24

11. 「Excel 出力」アイコンをダブルクリックしファイル名欄に Excel ファイルの出力先を入力してください。

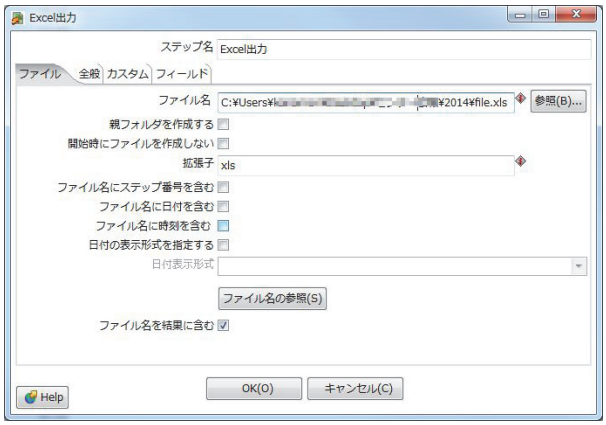

図 25

また「フィールド」タブをクリックし、「フィールドを取得」ボタンをクリックし、「OK」ボタンをクリックします。

図 26

12. 「実行」ボタン(図 27 参照)をクリックしてください。

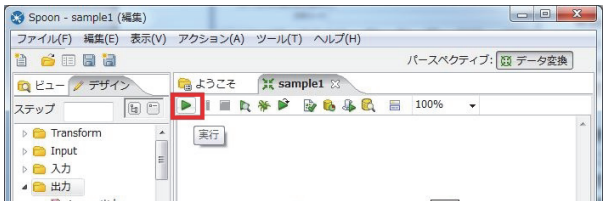

図 27

すると図 28 のような Excel ファイルが作成されます。

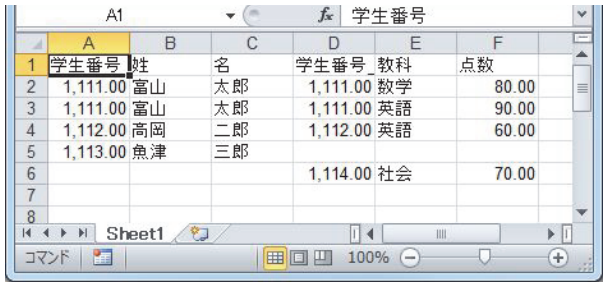

図 28

なお手順[9]の結合タイプ入力欄にて「INNER」を選択した場合、図 29 のような Excel になります。

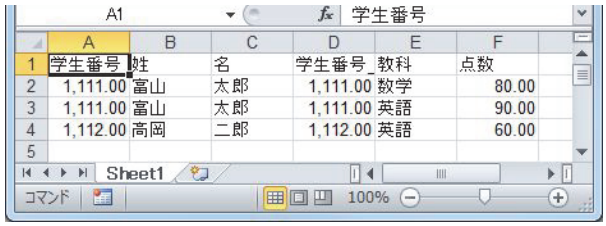

図 29

なお、学籍番号および点数が小数点第 2 位まで表示されていますが、Excel 入力時のデータタイプが”Number”となっているためです。データタイプを”String”にすると小数点表示は無くなります。

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。