SRA の技術情報誌
GS LetterNeo
Vol.171　2022 年 10 月

オープンソース ETL ツール Talend Open Studio の紹介

田中 克憲

アドバンストクラウドエンジニアリング事業部

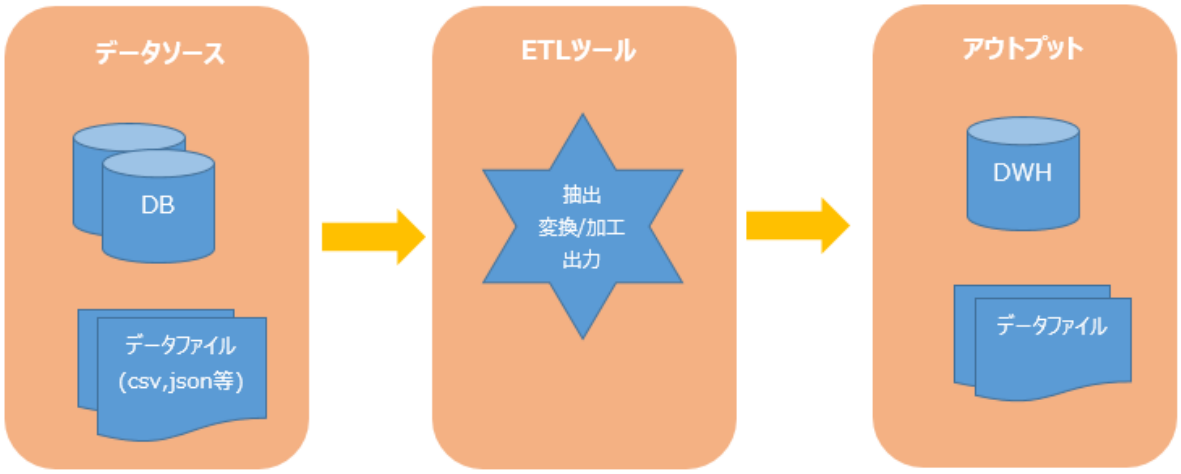
はじめに

データを収集・加工・結合する ETL ツールは様々な製品が世に出ています。それぞれに特色が異なっているので、ETL ツールの導入を検討する際にはどの製品を選定するか悩ましいところではありますが、本書ではオープンソースの ETL ツールである「Talend Open Studio[1]」を紹介します。

[1] https://www.talend.com/jp/products/talend-open-studio/

ETL ツールとは

ETL は、Extract(抽出), Transform(変換), Load(書き出し) のそれぞれの単語の頭文字を略したものです。あらゆるデータソースに蓄積されたデータを抽出し、用途に応じて利用しやすい形式に変換・加工した上で、データウェアハウスなどの格納先に書き出してくれるツールです。



● ETL ツールの必要性

ETL のプロセスである、抽出・変換・加工・出力 はプログラミングでも実現することができますが、ETL ツールを利用することで開発工数を削減できるメリットがあります。また、ETL ツールはプログラミングの知識が少なくても開発できるため高度な専門知識を有する人材を確保しなくても開発作業を行うことが可能です。

● ETL ツールの用途

- ✓ 散在するデータの取集・集約化
- ✓ データ更新作業の機械化・自動化
- ✓ 新システムへのデータ移行

などが挙げられます。

Talend Open Studio について

Talend Open Studio は、無償で利用できるオープンソース ETL ツールです。有償版の「Talend Data Fabric」もありますが、無償版の Talend Open Studio はトライアルではなく制限も多くないため、基本的な ETL 開発であれば十分に機能すると思います。標準で数多くのコンポーネント(処理部品)が用意されており、GUI 操作でコンポーネントを配置し、各コンポーネントを繋ぎ合わせることでデータ入力から変換、出力までのフローを構築することが可能です。

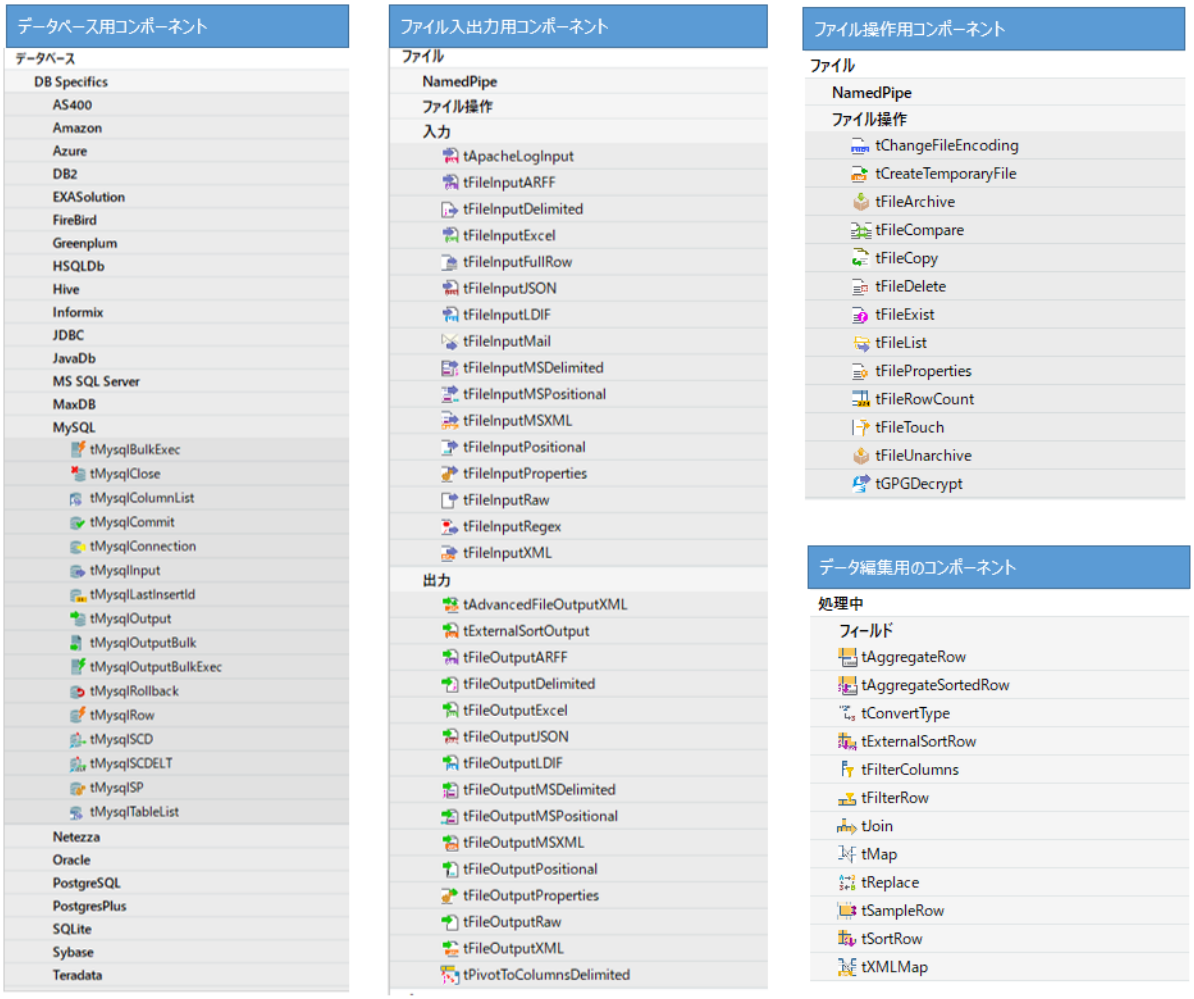


図 1 標準で用意されているコンポーネントの一例

以下の図 2 は Talend Open Studio の画面です。青枠の部分がコンポーネント一覧です。ここから使用したいコンポーネントを選択し、赤枠のデザインワークスペースに矢印の様にドラックアンドドロップすることでコンポーネントを配置していきます。図 2 は Oracle データベースからデータを入力するコンポーネントを配置した状態です。

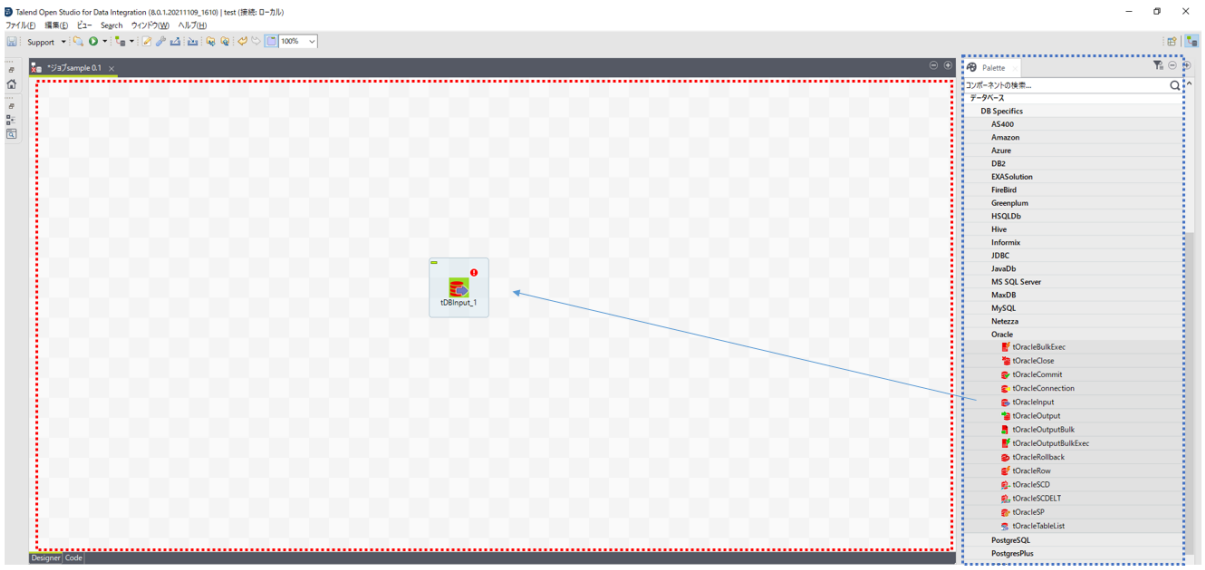


図 2　Talend Open Studio の画面

実際に動かしてみる

　簡単なサンプルを作成してみます。入力データとしてデータベース（MySQL）に以下の 2 テーブルを用意します。

■inputTBL_A

id	A_column1	A_column2
1	A_col1_data1	A_col2_data1
2	A_col1_data2	A_col2_data2

■inputTBL_B

id	B_column1	B_column2	B_column3	B_column4
1	B_col1_data1	B_col2_data1	B_col3_data1	B_col4_data1
2	B_col1_data2	B_col2_data2	B_col3_data2	B_col4_data2
3	B_col1_data3	B_col2_data3	B_col3_data3	B_col4_data3
4	B_col1_data4	B_col2_data4	B_col3_data4	B_col4_data4

Talend Open Studio を使用して、上記のデータを抽出/変換し、以下の形式で別テーブルに出力する仕組みを構築していきます。

■outputTBL_C

id	C_column1	C_column2	C_column3
1	A_col1_data1&A_col2_data1	B_col1_data1	B_col2_data1
2	A_col1_data2&A_col2_data2	B_col1_data2	B_col2_data2

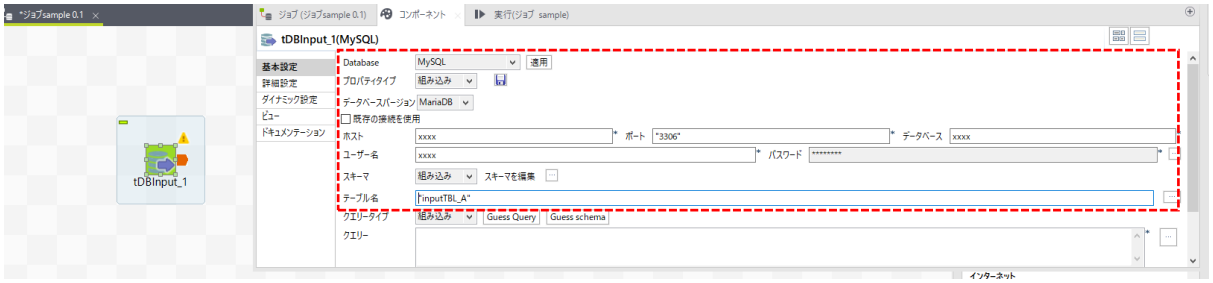
■outputTBL_D

id	D_column1	D_column2	D_column3
1	A_col2_data1	B_col3_data1	B_col4_data1
2	A_col2_data2	B_col3_data2	B_col4_data2

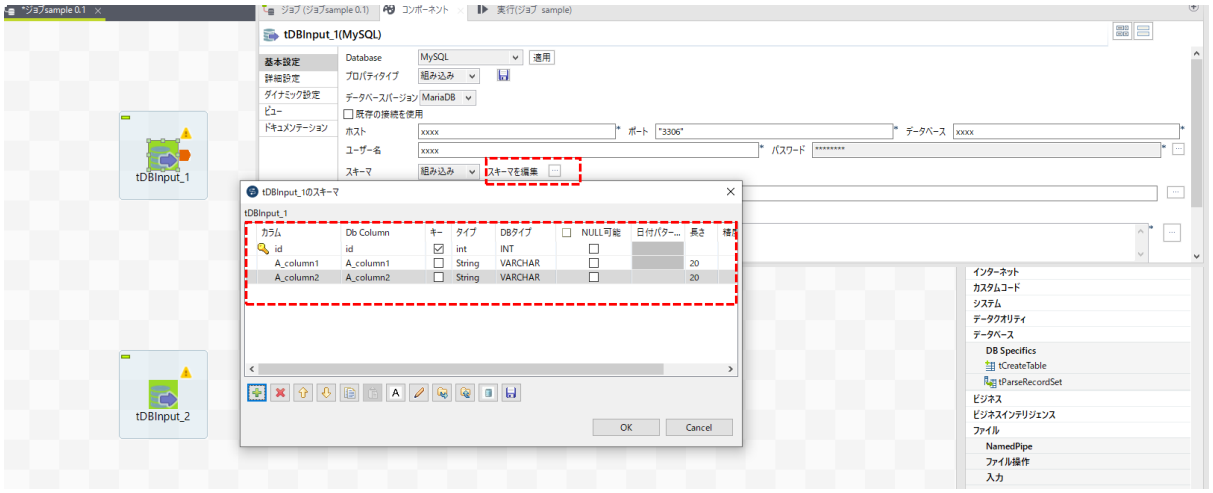
[STEP1： 入力用コンポーネントの配置]

MySQL テーブルからデータを抽出するコンポーネントを 2 つ配置します。

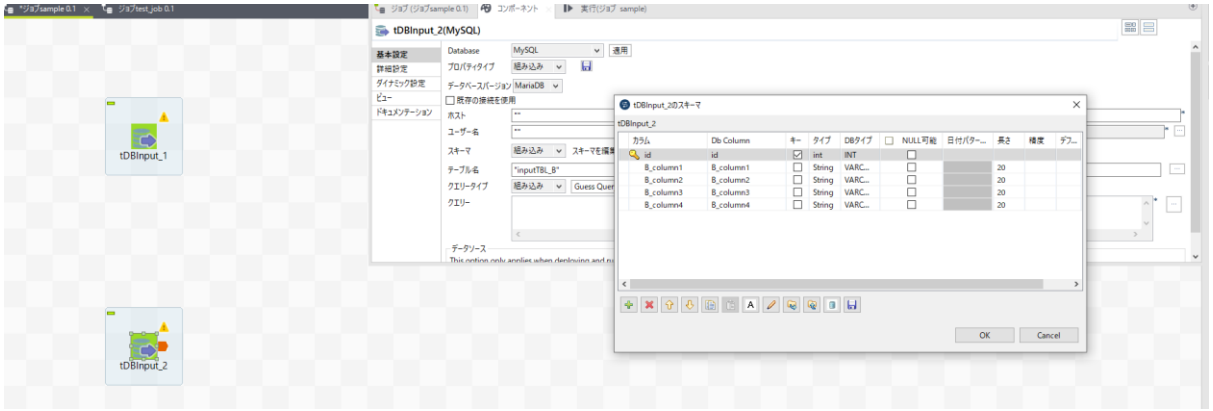
inputTBL_A テーブルのデータを抽出するために、コンポーネント(tDBInput_1)の設定を行います。ここではデータベース接続情報と抽出対象テーブル名を入力します。



次に抽出対象カラムを入力します。

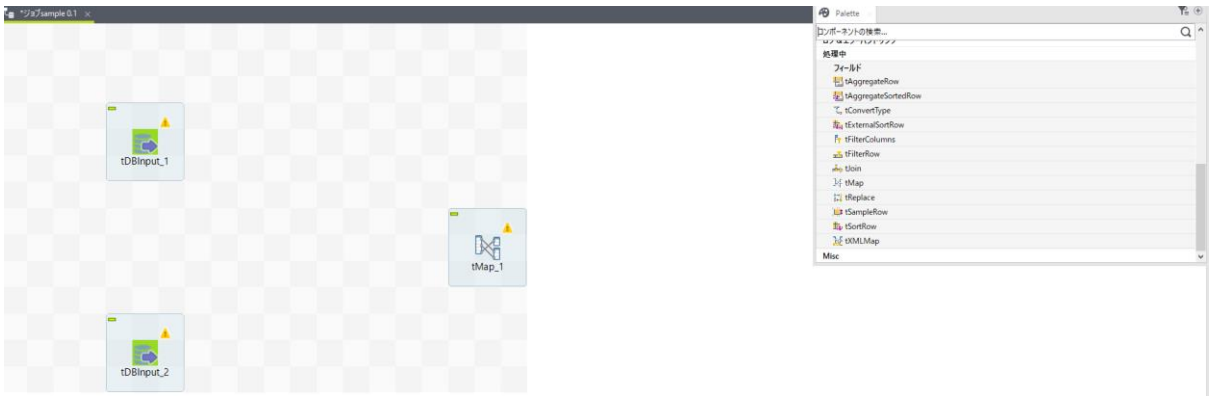


inputTBL_B テーブルのデータを抽出するために、同様の設定をコンポーネント(tDBInput_2)に行います。

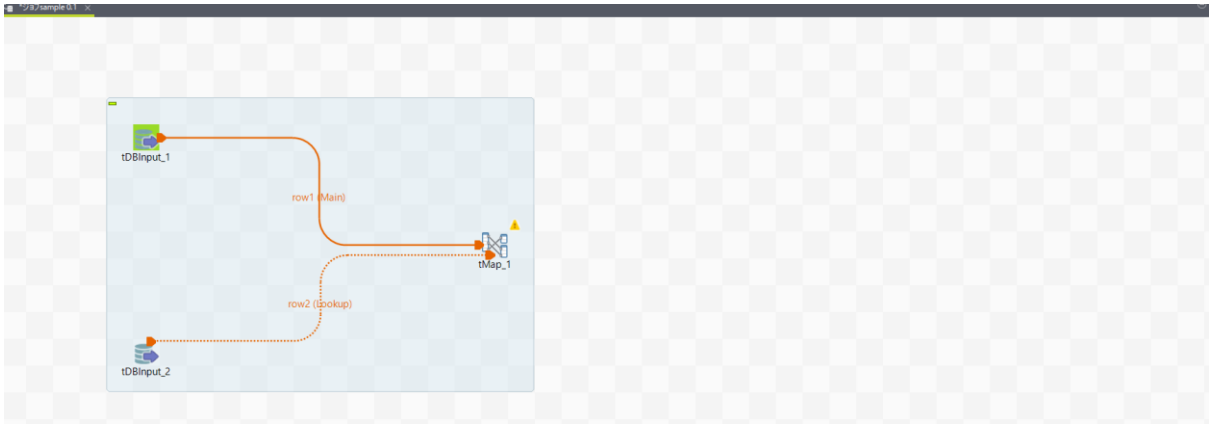


[STEP2： データ変換用コンポーネントの配置]

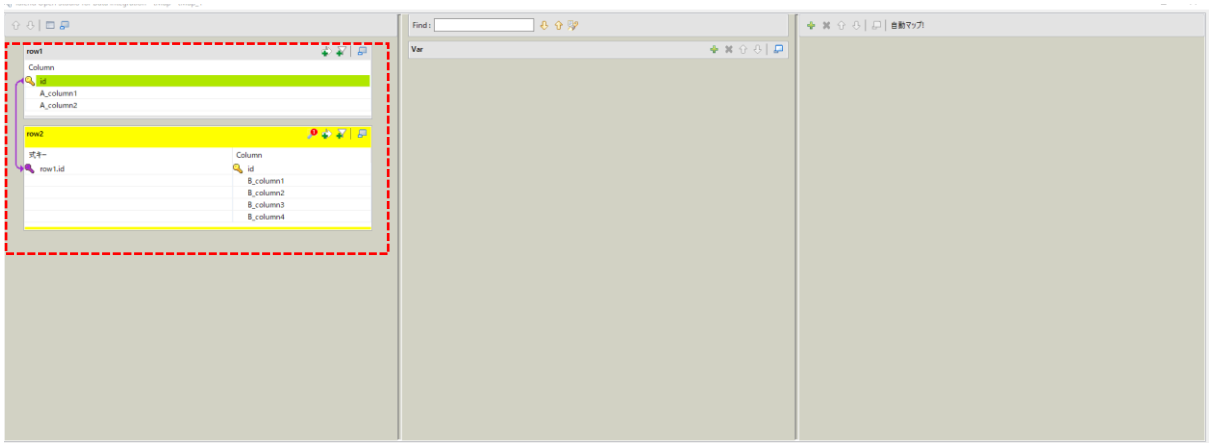
入力データを変換するためのコンポーネント(tMap)を配置します。Talend Open Studio を使用する上で、この tMap は使用頻度が高く最も重要なコンポーネントです。



tMap を配置後、tDBInput_1 と tDBInput_2 を tMap に繋ぎます。これで tMap へデータを受け渡すことが出来ます。

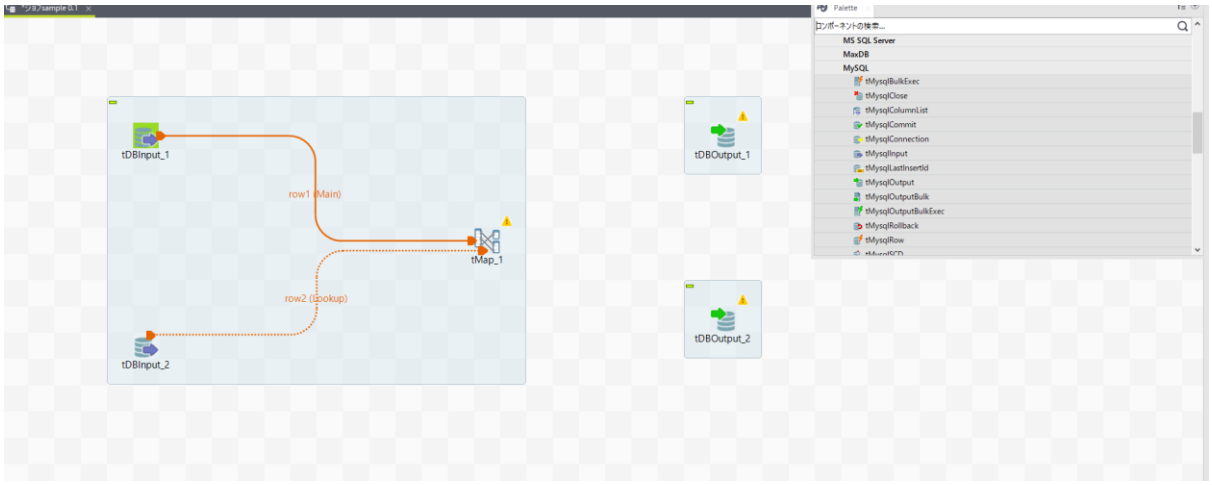


tMapの設定画面を開くと、tDBInput_1(inputTBL_A)とtDBInput_2(inputTBL_B)のカラムが表示されます。tMap は左側に入力先、右側に出力先を表示しますが、今の時点では出力先のコンポーネントを用意していないので右側には何も表示されていません。inputTBL_A と inputTBL_B を結合するため tMap 上で id を紐づけます。

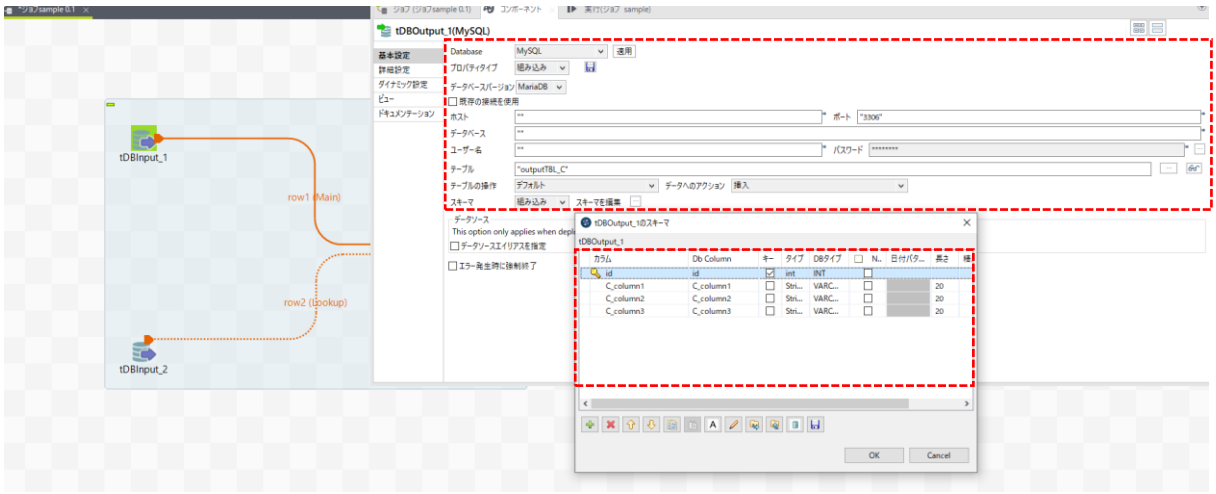


[STEP3： 出力用コンポーネントの配置]

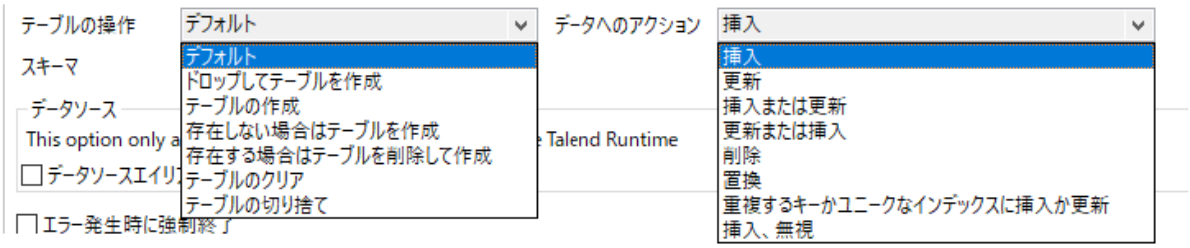
MySQL テーブルにデータを格納するコンポーネントを 2 つ配置します。



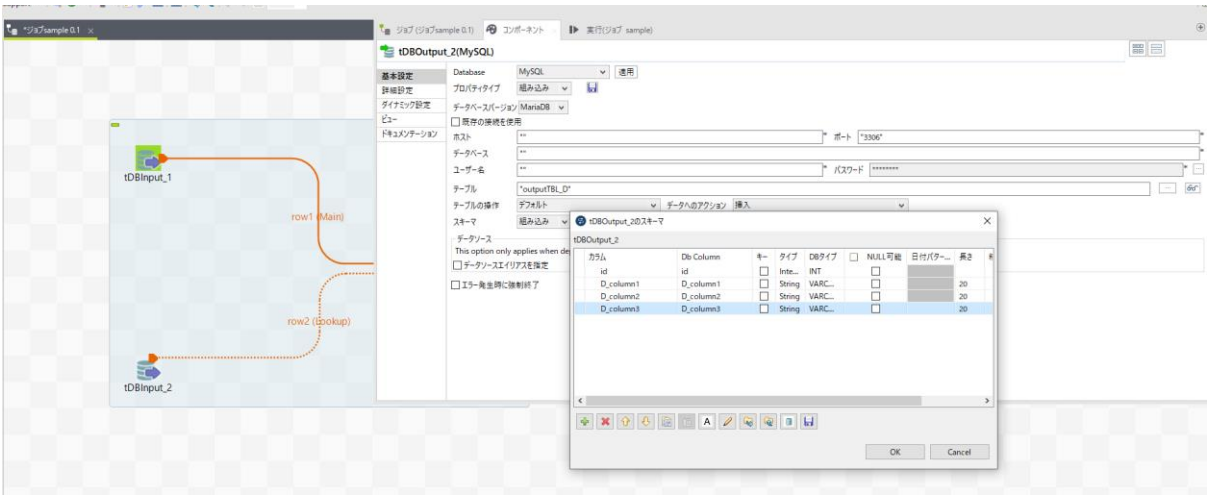
outputTBL_C テーブルにデータを格納するためのコンポーネント(tDBOutput_1)の設定を行います。入力用コンポーネントの設定と同じようにデータベース接続情報と出力対象テーブル名、対象カラムを入力します。



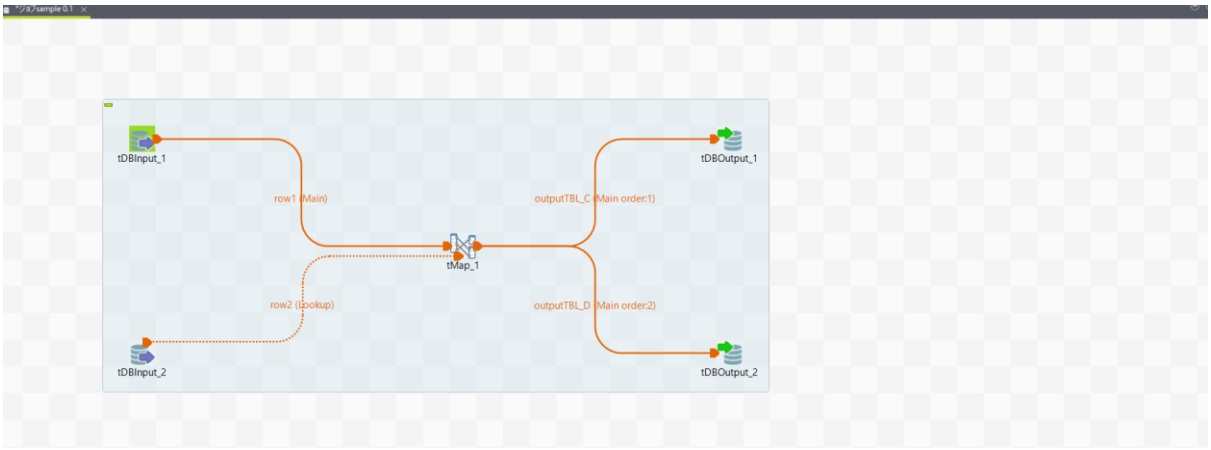
データ出力時のテーブル操作、データアクションを指定することも出来ます。



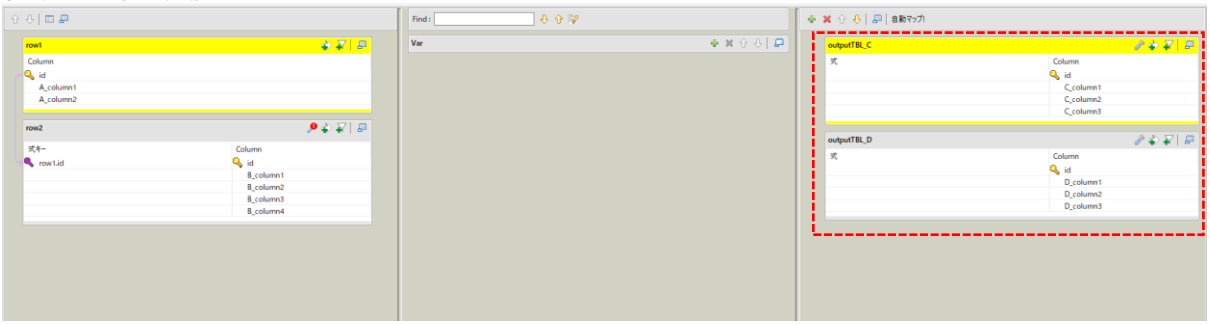
outputTBL_D テーブルにデータを格納するために、同様の設定をコンポーネント(tDBOutput_2)に行います。



設定が完了した tDBOutput_1 と tDBOutput_2 を STEP2 で配置した tMap に繋ぎます。

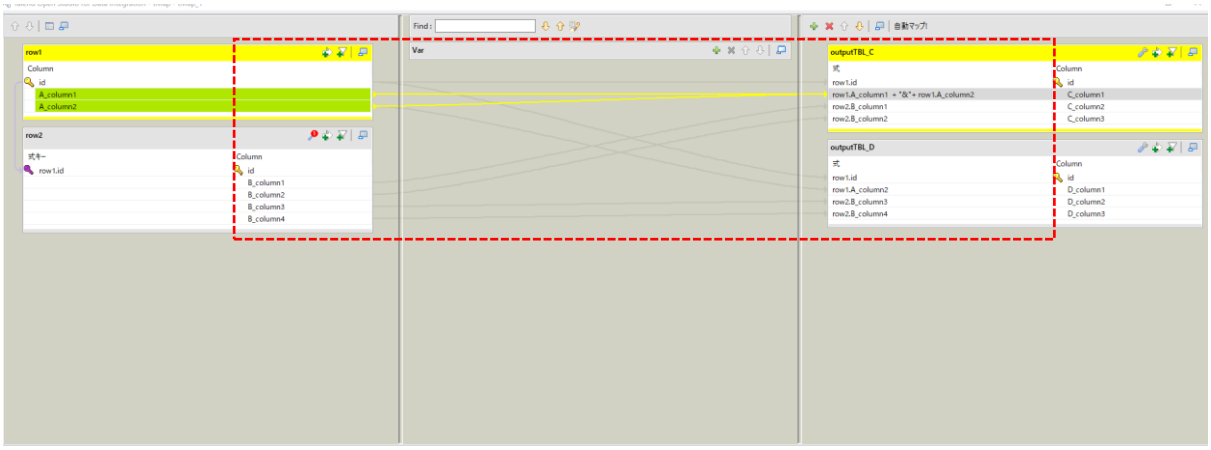


tMapの設定画面を開くと、tDBOutput_1(outputTBL_C)とtDBOutput_2(outputTBL_D)のカラムが右側に表示されます。



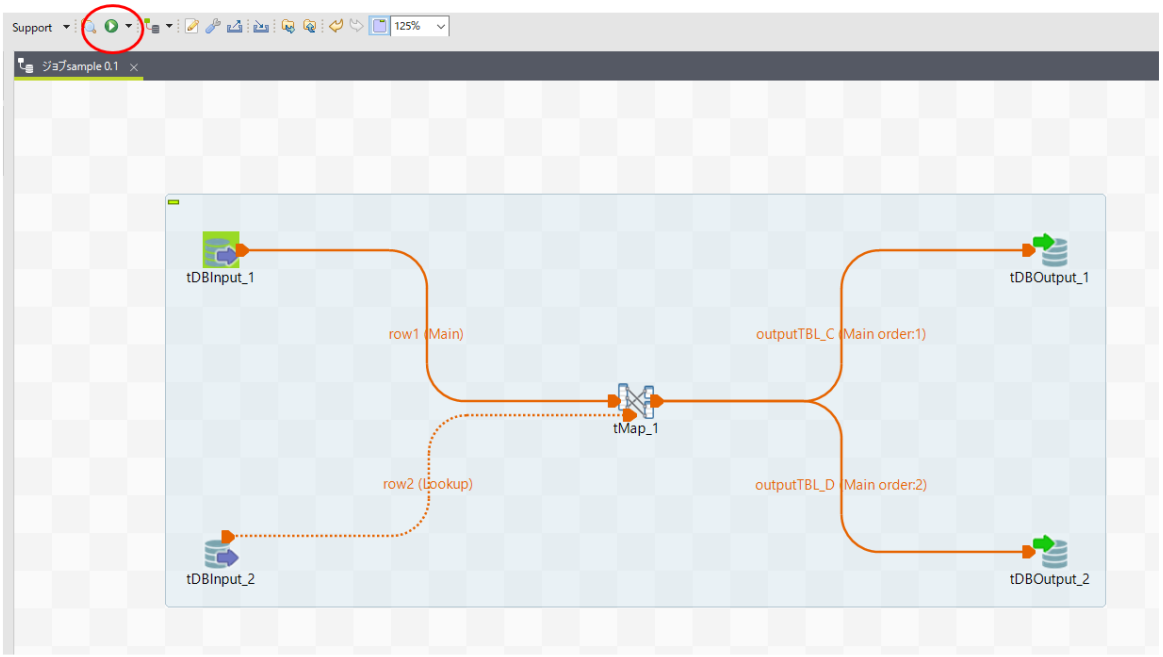
[STEP4： tMap の入出力設定]

データの入力元と出力先が揃ったところで、入出力の設定を行います。入力元のカラムをドラッグ&ドロップして出力先のカラムに紐づけます。

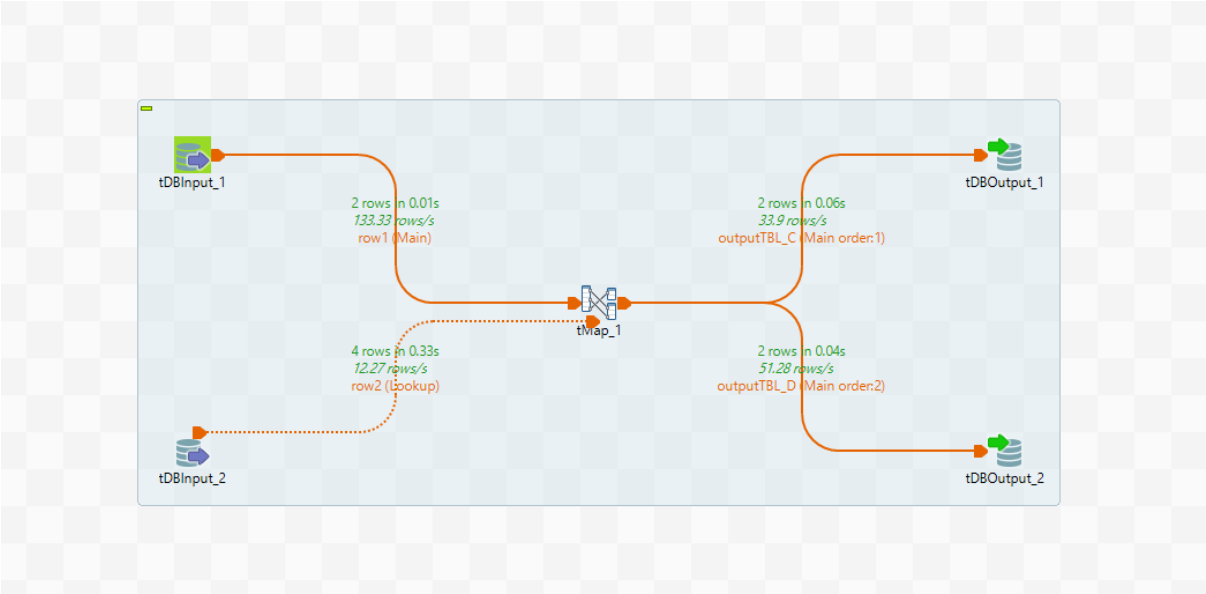


今回は tMap の式を使用した単純な変換処理として、A_column1 カラムと A_column2 カラムの値を文字列連結した値を outputTBL_C テーブルの C_column1 カラムに格納しています。変換処理を記述する式には Java の構文を使用することが出来るので「row1.kbn == 0?”Yes”:”No”」といった三項演算子を使用した変換も可能です。

以上でサンプルの作成は完了です。それでは処理を実行してみましょう。赤枠の実行ボタンを押下します。



処理が正常終了すると、以下のように処理結果が緑色で表示されます。



出力先のテーブルデータを確認します。

■outputTBL_C

id	C_column1	C_column2	C_column3
1	A_col1_data1&A_col2_data1	B_col1_data1	B_col2_data1
2	A_col1_data2&A_col2_data2	B_col1_data2	B_col2_data2

■outputTBL_D

id	D_column1	D_column2	D_column3
1	A_col2_data1	B_col3_data1	B_col4_data1
2	A_col2_data2	B_col3_data2	B_col4_data2

5 ページに記載した以下の想定通りにデータを出力することができました。

■outputTBL_C

id	C_column1	C_column2	C_column3
1	A_col1_data1&A_col2_data1	B_col1_data1	B_col2_data1
2	A_col1_data2&A_col2_data2	B_col1_data2	B_col2_data2

■outputTBL_D

id	D_column1	D_column2	D_column3
1	A_col2_data1	B_col3_data1	B_col4_data1
2	A_col2_data2	B_col3_data2	B_col4_data2

おわりに

今回は簡単なサンプルによる Talend Open Studio の紹介になりましたが、様々なコンポーネントを使用することで、より複雑な変換・加工処理を行うことができます。また、ファイル入出力コンポーネントも豊富に備わっているので CSV ファイル、JSON ファイルといったファイルデータとデータベースのデータを組み合わせて変換・加工することもできます。ETL ツールの導入を検討する際に Talend Open Studio も選択肢に加えてみてはいかがでしょうか。

参考 URL

https://www.talend.com/jp/products/talend-open-studio/

Talend Open Studio for Data Integration 入門ガイド
https://help.talend.com/r/ja-JP/8.0/studio-getting-started-guide-open-studio-for-data-integration/introduction

Talend Open Studio for Data Integration User Guide
https://help.talend.com/r/ja-JP/8.0/studio-user-guide-open-studio-for-data-integration

GSLetterNeo Vol.171
2022 年 10 月 20 日発行
発行者　株式会社 SRA 技術本部 先端技術研究室

編集者　熊澤努　方 学芬
バックナンバー https://www.sra.co.jp/public/sra/gsletter/
お問い合わせ　gsneo@sra.co.jp

株式会社SRA
〒171-8513 東京都豊島区南池袋 2-32-8

夢を。Yawaraka Innovation
やわらかいのべーしょん