Lab9: Data Engineering [Solutions]

Extraction Process:

1. Collect and create Metadata for new movie dataset.

Transform Process:

1. Create Job5CleanMovieBudget and add 3 components including tInputDelimited(Metadata of new movie) , tMap , tLogRow

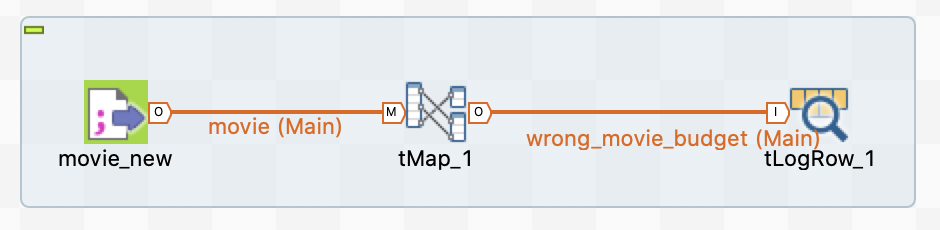


Fig 1, Job design of Job5CleanMovieBudget

1. For Job5CleanMovieBudget, clean movie’s budgets by creating variable named “chk\_digit” and using the function Mathematical.NUM(<var>) to check digit value of movie’s budgets.

**Note**: Mathematical.NUM(<var>) will return 1 if the string parameter is all digits.

[Task1] Capture screenshot of the expression builder of variable “chk\_digit”

Solution:

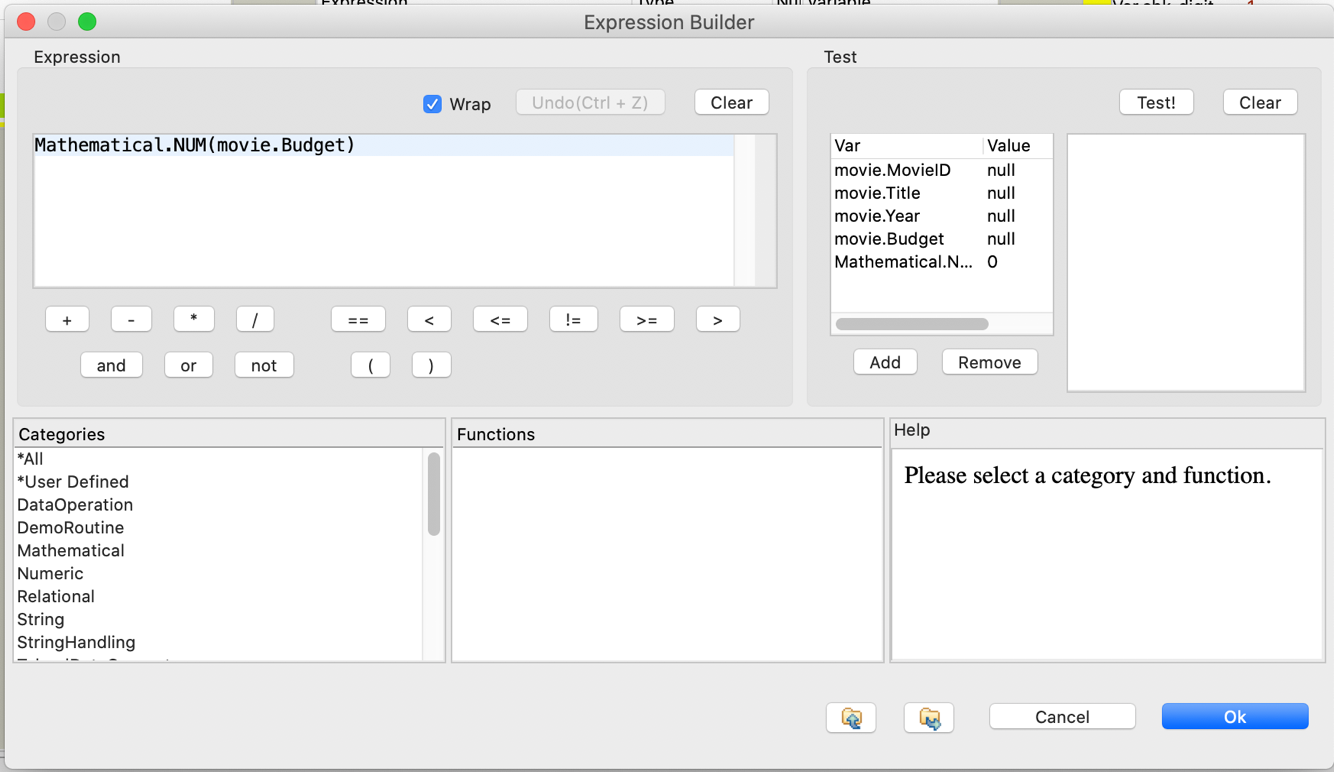


Fig 2., the expression builder of variable “chk\_digit”

1. Create 2 output tables which comprise four columns : MovieID, Title, Year, Budget.
   1. Create output table#1 named “cleaned\_movie\_budget” => Filter the movie’s budgets that are numeric values
   2. Create output table#2 named “wrong\_movie\_budget” => Filter the movie’s budgets that are not numeric values

[Task2] Capture screen of mapping process in tMap which show input table, output table and variable

Solution:

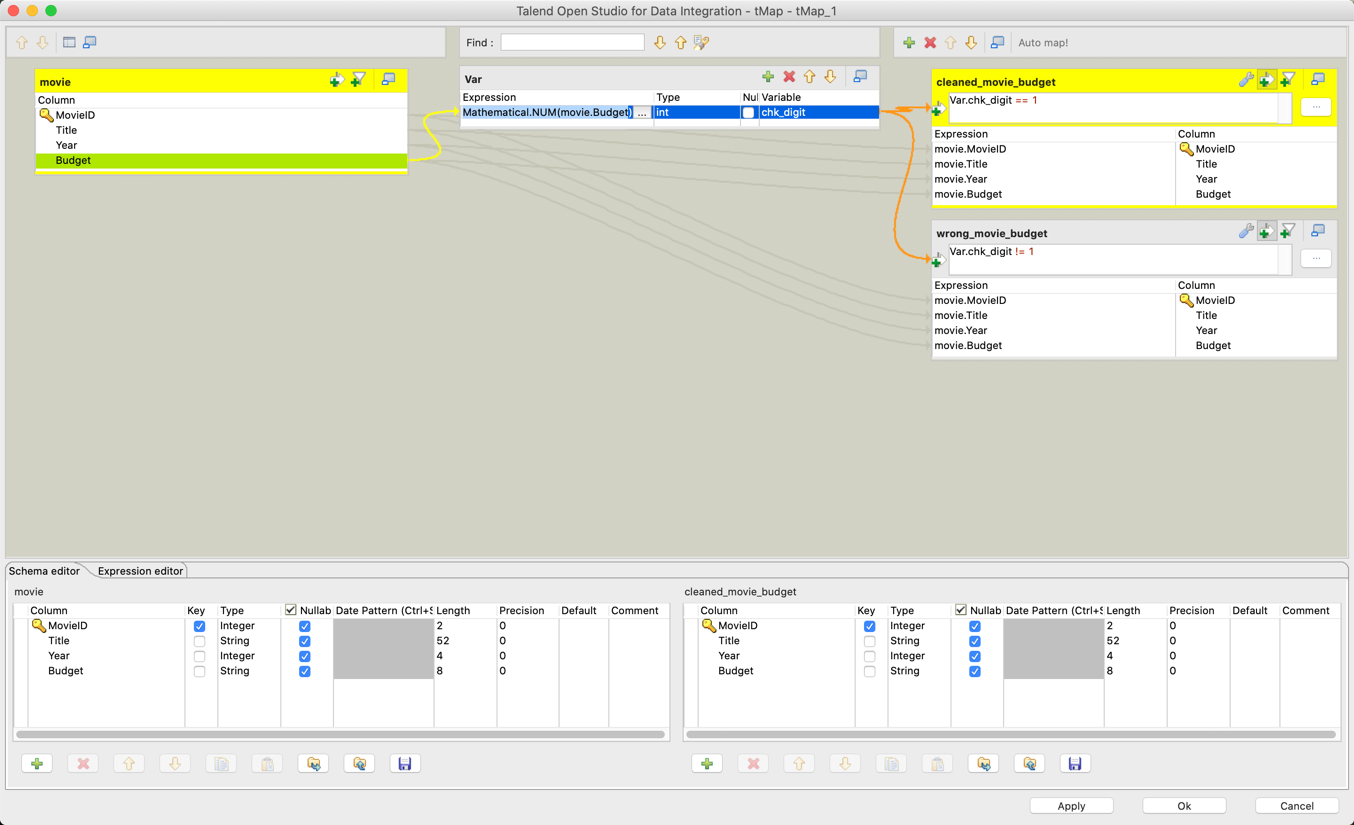


Fig 3., mapping process in tMap

[Task3] Use tLogRow to print the output of table “wrong\_movie\_budget” and capture screenshot of the output.

Solution:

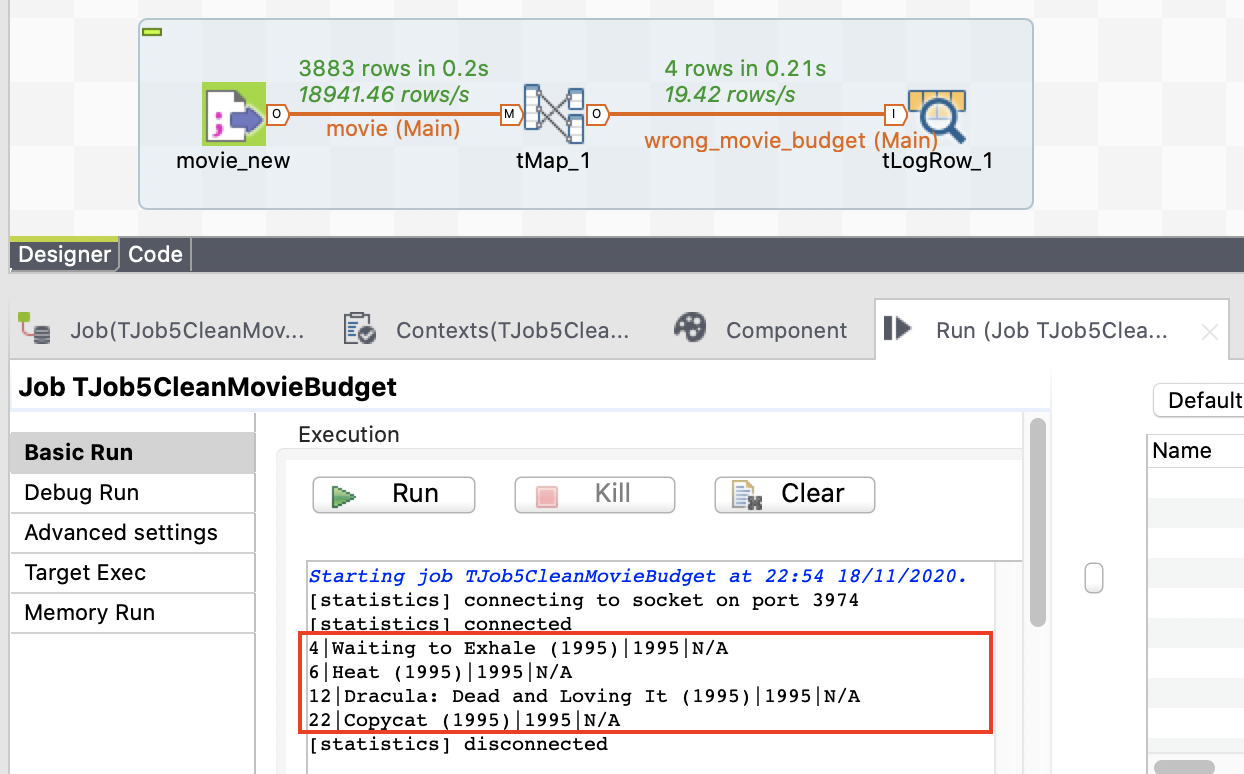


Fig 4., output of table “wrong\_movie\_budget”

1. Create “Job6ETLMovieRatingNew” by duplicating the job from “TJob4ETLMovieRating” (Righ click at the TJob4ETLMovieRating and select Duplicate )
2. Replace metadata of “movie” with the input flow of cleaned movie data from “Job5CleanMovieBudget” in the new job “Job6ETLMovieRatingNew” as shown in Fig 3 and Fig 4.

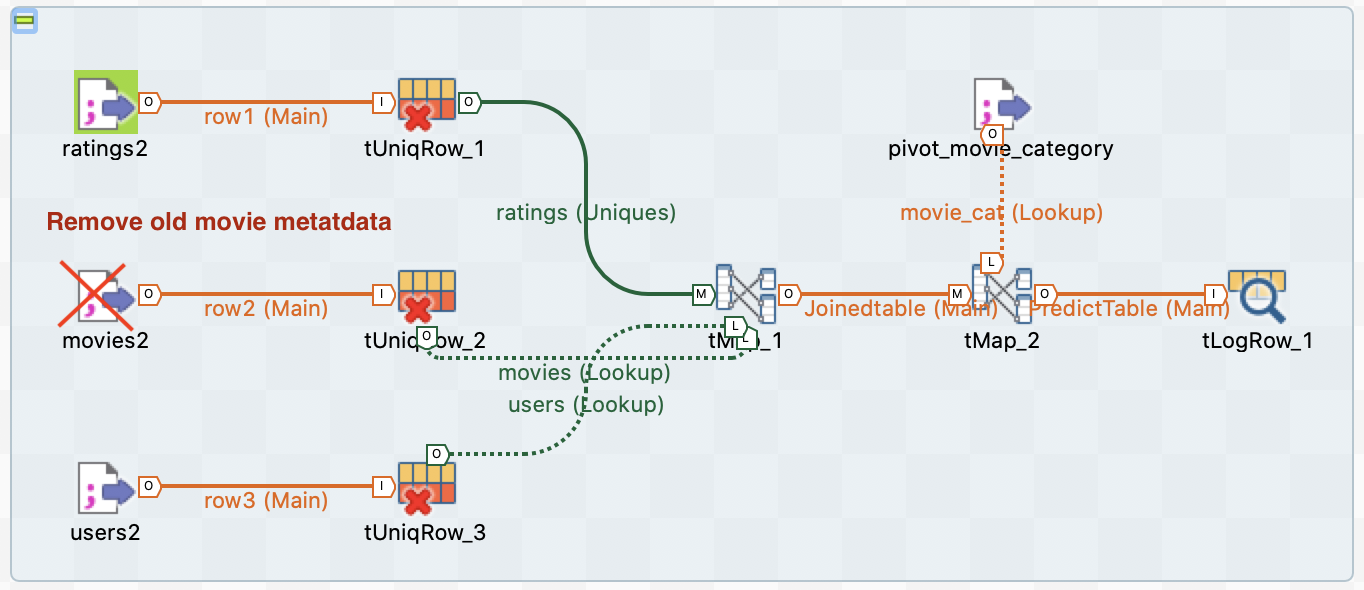


Fig 5, Job design of Job4ETLMovieRating

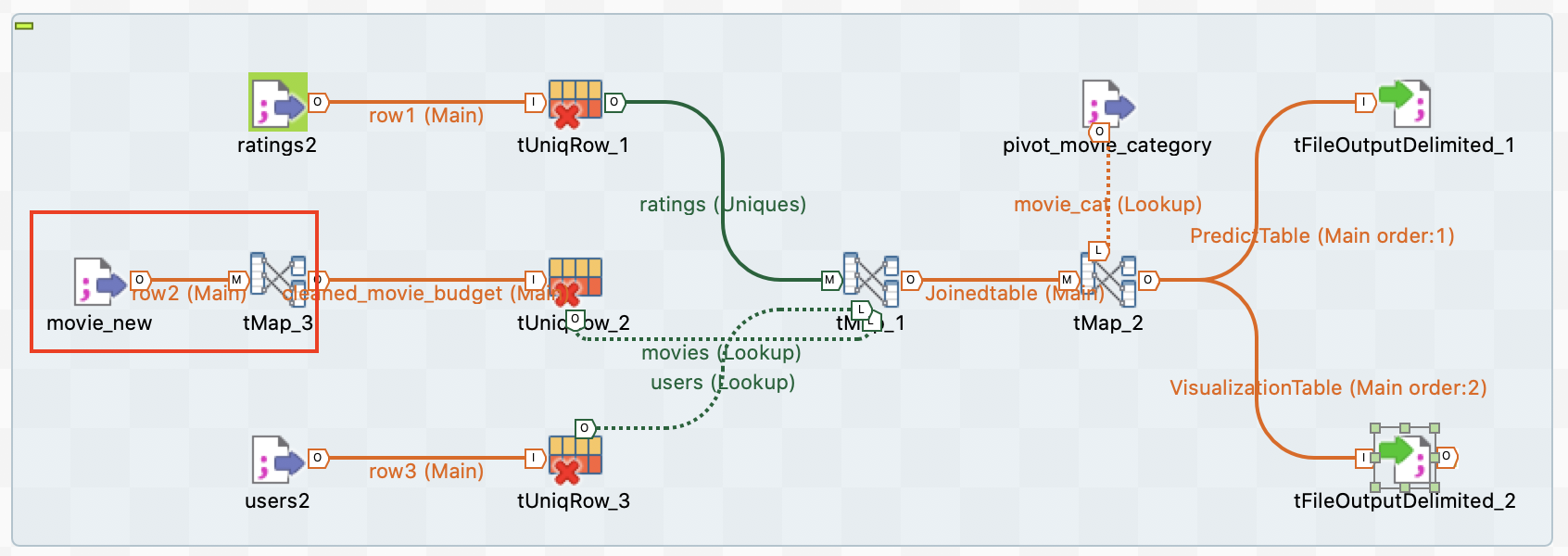
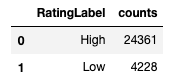


Fig 6, Job design of Job6ETLMovieRatingNew

1. Add Budget column into the output of PredictTable.
2. Add a new column named “RatingLabel” and transform rating data to binary values (use IF/ELSE Statement).
   1. Rating value >=3 is “High”
   2. Rating value <3 is “Low”

[Task4] Count the number of rating labels of “High” and “Low”. (2 answers)

**Solution**:



1. The number of “High” label is 24361
2. The number of “Low” label is 4228