Lab9: Data Engineering

Objectives: Learn ETL Process using Talend Open Studio for Data Integration about collecting and extracting data from multiple sources, transforming the data to the desired format, and write the output to the target datasource.

Estimated Time : 2 hours

Lab Instruction

The tasks of this assignment is continue from a talend project in the lab class. In this assignment, we need to add the movie's budget as one of the feature to predict the movie rating. The features to use in the prediction model are age, gender, occupation, movie category ,and movie budget. There are four columns in a new movie dataset including MovieID, Title, Year and Budget. Therefore, we need to create new metadata of movie and merge with the user profile and movie rating datasets. In the transform process, we need to clean the budget values by separating out the unknown movie's budget (N/A) before using in the prediction model.

MovieID	Title	Year	Budget
1	Toy Story (1995)	1995	10685425
2	Jumanji (1995)	1995	18656746
3	Grumpier Old Men (1995)	1995	7323119
4	Waiting to Exhale (1995)	1995	N/A
5	Father of the Bride Part II (1995)	1995	10777966
6	Heat (1995)	1995	N/A
7	Sabrina (1995)	1995	9498194
8	Tom and Huck (1995)	1995	11636471
9	Sudden Death (1995)	1995	14600147
10	GoldenEye (1995)	1995	19183206
11	American President, The (1995)	1995	17447917
12	Dracula: Dead and Loving It (1995)	1995	N/A

Fig1, movie_new.csv

Extraction Process:

1. Collect and create Metadata for new movie dataset.

Transform Process:

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

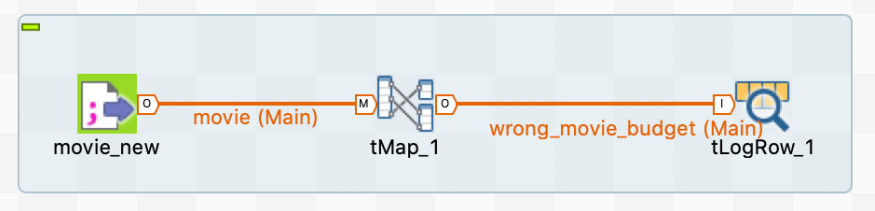

Fig 2, Job design of Job5CleanMovieBudget

3. 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"

4. Create 2 output tables which comprise four columns : MovieID, Title, Year, Budget.
   a. Create output table#1 named "cleaned_movie_budget" => Filter the movie's budgets that are numeric values
   b. 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

[Task3] Use tLogRow to print the output of table "wrong_movie_budget" and capture screenshot of the output.

5. Create "Job6ETLMovieRatingNew" by duplicating the job from "TJob4ETLMovieRating" (Righ click at the TJob4ETLMovieRating and select Duplicate)
6. 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.

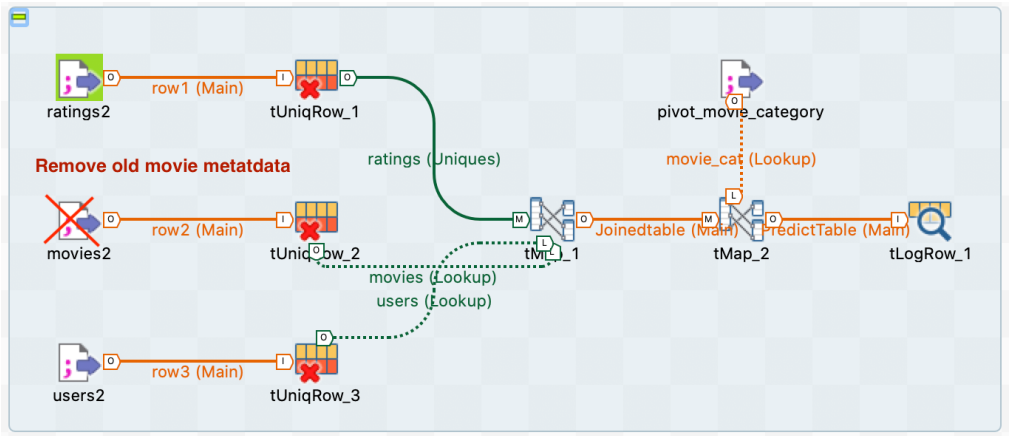

Fig 3, Job design of Job4ETLMovieRating

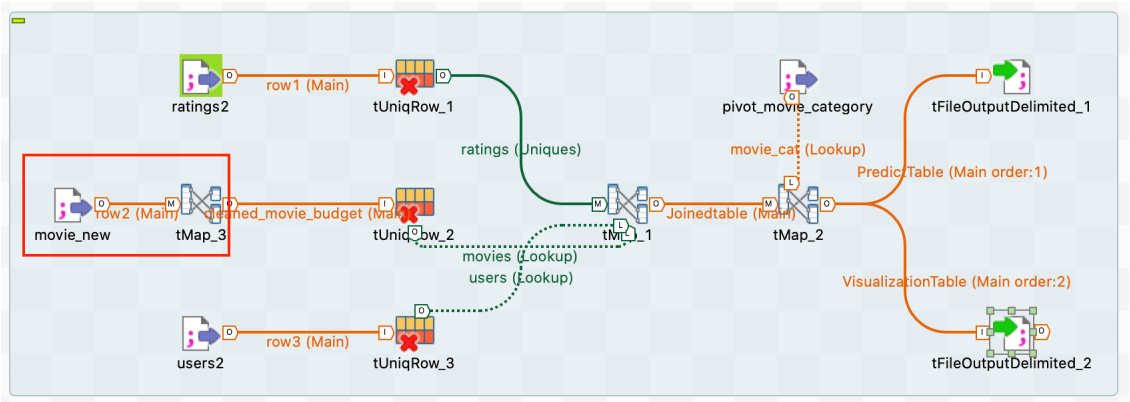

Fig 4, Job design of Job6ETLMovieRatingNew

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

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

Lab Submission

Submission System: Google Classroom

Total TASKS: 4