Lab9: Data Engineering [Solutions]

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

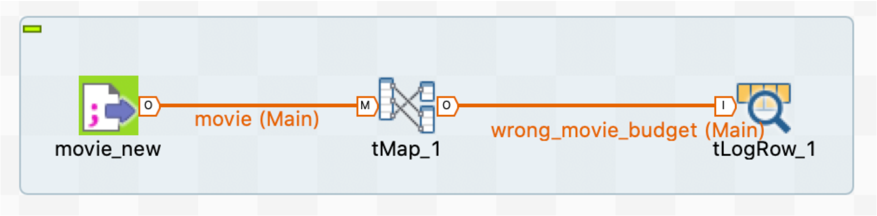
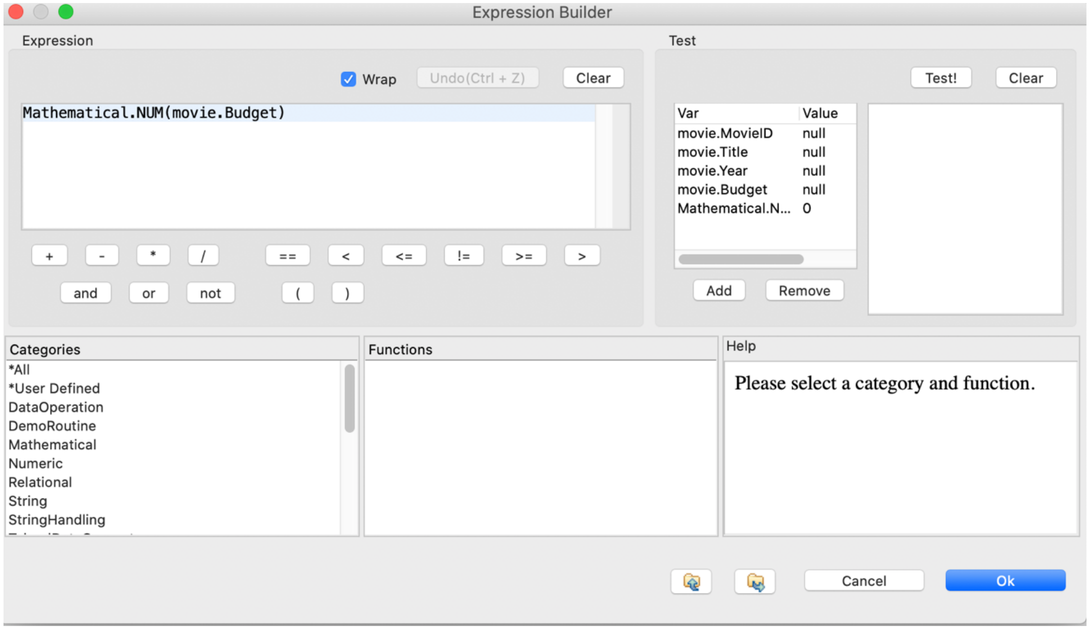


Fig 1, Job design of Job5CleanMovieBudget

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"

Solution:



```
Mathematical.NUM(movie.Budget)
```

Var	Value
movie.MovieID	null
movie.Title	null
movie.Year	null
movie.Budget	null
Mathematical.N...	0

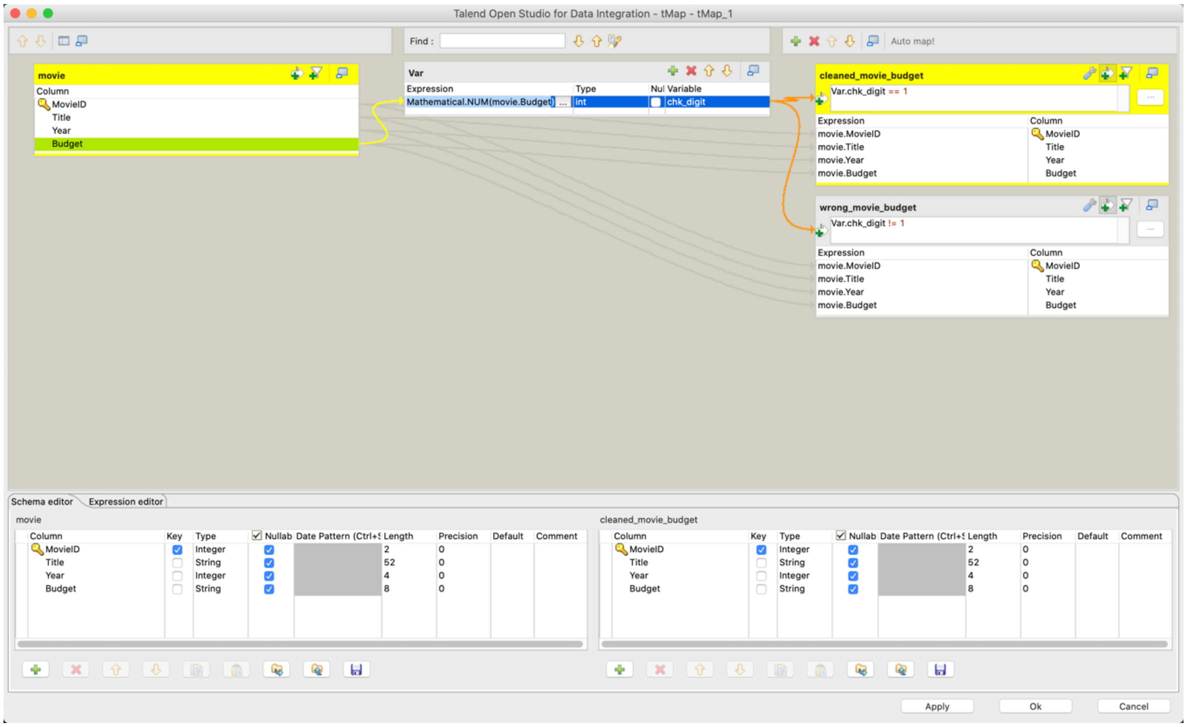
Please select a category and function.

Fig 2., 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

Solution:



movie

Column
MovieID
Title
Year
Budget

Var

Expression	Type	Nul	Variable
Mathematical.NUM(movie.Budget)	int	□	chk_digit

cleaned_movie_budget

Var.chk_digit == 1

Expression	Column
movie.MovieID	MovieID
movie.Title	Title
movie.Year	Year
movie.Budget	Budget

wrong_movie_budget

Var.chk_digit != 1

Expression	Column
movie.MovieID	MovieID
movie.Title	Title
movie.Year	Year
movie.Budget	Budget

movie

Column	Key	Type	☑ Nullab	Date Pattern (Ctrl+S	Length	Precision	Default	Comment
MovieID	☑	Integer	☑		2	0		
Title	□	String	☑		52	0		
Year	□	Integer	☑		4	0		
Budget	□	String	☑		8	0		

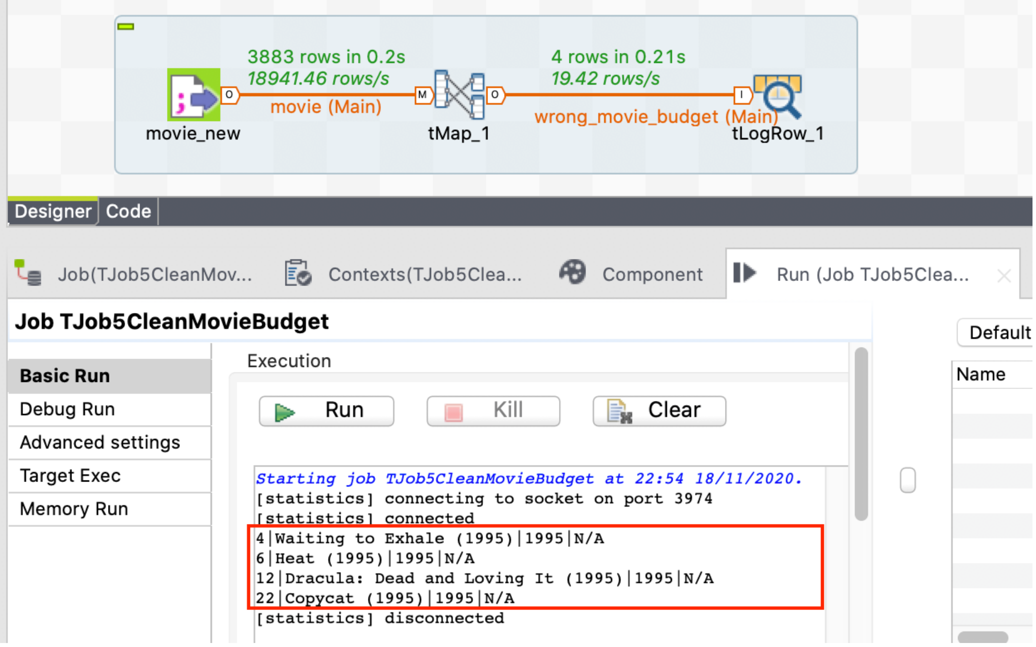
cleaned_movie_budget

Column	Key	Type	☑ Nullab	Date Pattern (Ctrl+S	Length	Precision	Default	Comment
MovieID	☑	Integer	☑		2	0		
Title	□	String	☑		52	0		
Year	□	Integer	☑		4	0		
Budget	□	String	☑		8	0		

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:



```
Starting job TJob5CleanMovieBudget at 22:54 18/11/2020.
[statistics] connecting to socket on port 3974
[statistics] connected
4|Waiting to Exhale (1995)|1995|N/A
6|Heat (1995)|1995|N/A
12|Dracula: Dead and Loving It (1995)|1995|N/A
22|Copycat (1995)|1995|N/A
[statistics] disconnected
```

Fig 4., output of table “wrong_movie_budget”

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.

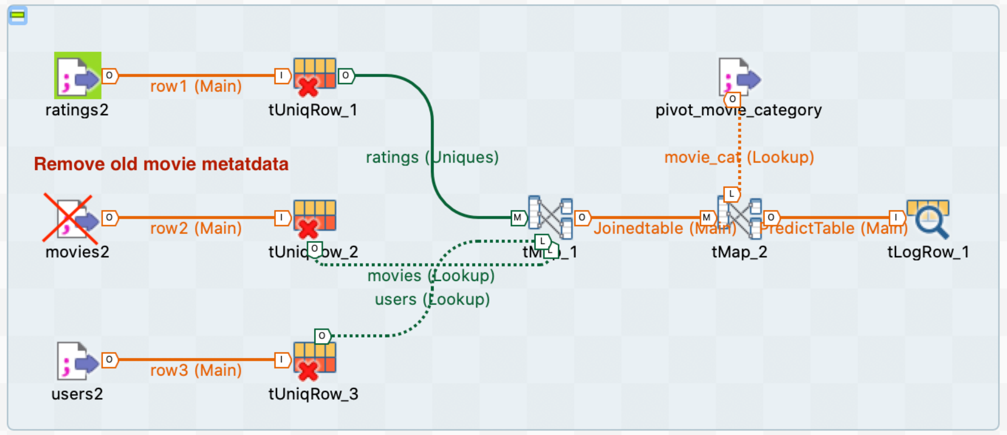


Fig 5, Job design of Job4ETLMovieRating

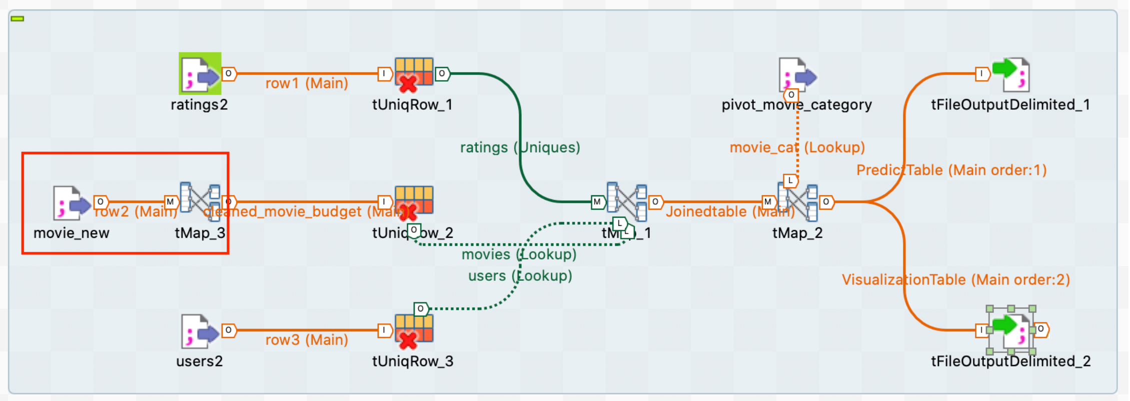
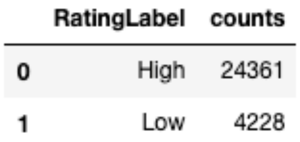


Fig 6, Job design of Job6ETLMovieRatingNew

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)

Solution:



	RatingLabel	counts
0	High	24361
1	Low	4228

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