Name: _____ Due: _____/_____/_____

R·I·T	**Rochester Institute of Technology** **Golisano College of Computing and Information Sciences** **School of Information**

Data Warehousing PE06
ETL Using Kettle

Overview

In this exercise you will have an opportunity to use Pentaho Data Integration (Kettle) to clean and transform input data

Scenario

PRODUCT_CATEGORY.csv

- There are duplicate categories. Strategy and Strat. should be the same.

PRODUCT.csv

- If the unit price is empty, then the price is $30
- If the unite price is negative, then the price needs to be changed back to positive
- Arthur Saves the Planet is an educational game

STEP:

1. For PRODUCT, set the unit price to $30 if the unit price is empty. (hint: use step **If field is null**)
2. For PRODUCT, set the unit price back to positive using Javascript.
3. For PRODUCT, sort rows on the Product according to categoryID. (hint: use step **Sort rows**)
4. For PRODUCT_CATEGORY, rename Strat. back to Strategy. (hint: use step **Replace in string**)
5. For PRODUCT_CATEGORY, sort rows on the Product Category according to categoryID.
6. Merge Join PRODUCT and PRODUCT_CATEGORY. (hint: use step **Merge Join**)
7. Delete the two categoryID columns. (hint: use step **Select values**)
8. Set the category for Arthur Saves the Planet to Education.
9. Use Table Output step to export cleaned product data to product table in PE_Kettle database.
   (Note: Create a MySQL database: PE_Kettle using the script file: pe_kettle.sql)
10. Place your solution (PE_KETTLE_yourname.ktr) into the drop box, and bring a hard copy of the grade sheet to in-person class on Monday, 11/2/20.

PRODUCT_CATEGORY.csv:

```
categoryID,categoryName
1,Education
2,Strategy
3,Strat.
4,Entertainment
5,Race
```

PRODUCT.csv:

```
ID,Name,unitPrice,categoryID
123456,Brain Bang 2.0,10,1
678901,Battleship,20,2
456789,Pictionary,10,4
234677,Risk,20,3
122565,Monopoly,,2
134144,Chess,-10,2
145723,Backgammon,-15,5
157302,Arthur Saves the Planet,6,
```

```
/* PE_Kettle.sql */
DROP DATABASE IF EXISTS PE_KETTLE;
CREATE DATABASE PE_KETTLE;
USE PE_KETTLE;
DROP TABLE IF EXISTS product;
CREATE TABLE product (
  id char(6) PRIMARY KEY,
  name varchar(30),
  unitPrice int,
  categoryName varchar(20));
```

Name: _____

R·I·T	**Rochester Institute of Technology** **Golisano College of Computing and Information Sciences** **School of Information**

PE06: ETL Using Kettle

Requirements	**Point Value**	**Points Earned**
Cleansing:		
- Remove duplication (Strategy and Strat.)	10	
- Change the unit price to $30 if the unit price is empty	10	
- Change the unit price back to positive if the unit price is negative	10	
- Change the category of Arthur Saves the Planet to education	10	
Transforming & loading:		
- All changes are made in Kettle	20	
- The kettle file can be run	20	
- The output must be loaded successfully to PE_Kettle database	20	
Points Earned	100	
Graded by		

Comments: