Pentaho data integration homework– beginners' course by Steinberg itamar

To start the homework you first need to:

1. Create a new database called: sakila_wh_homework as a clone of sakila_wh
2. Create a new connection in PDI to the new database.

Dim_time:

1. Create the full ETL of dim_time by yourself.

2. Name different steps that can replace the steps: "filter rows"," Add constant AM", "Add constant PM"

3. Use one of the steps from last question and change the ETL.

4. The manager of sakila asked that you will develop an ETL that creates a "targets" csv file.

   format
   the format of the file :
   "years", "months","staff_id", "rentals_count_2005", "target_rental_count_2006".

   columns:
   "years"
   hold a value of 2006 in format yyyy for each row.

   "months"
   Hold the 12 months of 2006 (01,02,03…)

   staff_id:
   the relevant staff id.

   rentals_count_2005:
   a column of real count of rentals of the corresponding year-month

   target_rental_count_2006:
   a column of real count * 1.1 as target. It should be round numbers (integers)

   The manager wants to see a row for each month even if there were no rentals for the specific staff id at that month. (hint: use cartesian step)

   Months that are null will get 100 as default target

years	months	Staff_id	rentals_count_2005	target_rental_count_2006
2006	01	1	0	100
2006	02	1	0	100

…				
2006	05	1	556	611
2006	05	2	600	660

You can use the “Merge Join”

```
select
year(rental_date) as years,month(rental_date) as months, staff_id, count(rental_id)
from
rental
where rental_date>='2005-01-01 00:00:00' and rental_date<='2005-12-31 23:59:59'
group by 1,2,3
```

Dim_date:

1. Create the full ETL of dim_date by yourself.
2. Add a columns to the stream. “Year-month” in the format of yyyy-mm.
3. Create the date_key without using the concat step.
4. Load data from a file into a table: “dim_date_from_file”

Dim_staff:

1. Create the full ETL of dim_staff by yourself.
2. Suggest a way to eliminate the “data grid” and “table output 2” – meaning add the -1 directly in the stream
3. Use a new step called “split fields” to extract the name of the staff member from the email column. Then find a way to eliminate the dot between the first name and last name. call the new field – “staff full name”
4. Which other steps can you use for task 3? java script – index of

Dim_store:

1. Create the full ETL of dim_store by yourself.
2. You have a file called: country_zipcode.csv, bring the zipcode column to the stream and add it to the dim_store you will need to learn about (stream lookup)

dim film:

1. think of a way to create the features 4 columns in a different way. Change the ETL to make it work.

2. the scenario is changed and now we have unknown number of features to each film. Think of 2 ways to handle such a case. Implement one of them

3. bring the release_year to the target

4. when do you think you need to use database lookup and when merge join? what are the advantages of database lookup?

5. is there a scenario you can think of that “merge join” will be more suited to the task

fact rentals:

1. use steps (not java script) in order to calculate rental_hours and add is_return.

2. use a step you know to turn full date to yyyymmdd instead of “string cut” and “Concat field”

3. instead of insert update change the ETL so it will bring everything each time you run the ETL