Module 5 Assignment
Data Integration using Pentaho

The Module 5 assignment provides experience with data integration tasks using Pentaho Data Integration. You need to perform some data cleaning operations on two data sources and then load the data into the fact table of the inventory data warehouse. You should see the class website for details about downloading Pentaho.

You need to create the inventory database tables before working on this assignment. Module 5 contains files with CREATE and INSERT statements for Oracle and PostgreSQL. You should create the inventory fact tables then execute the insert statements to populate the tables before starting on the transformation. Note that the *Currency_Dim* table is an additional table not part of the Inventory Data Warehouse design.

1. Data Sources

You need to perform cleaning operations on two data sources. Both data sources provide facts for the *Inventory_Fact* table of the inventory data warehouse. You need to create a transform for each data source.

Data Source 1: Access table

- Customer vendor key: integer
- Item master key: integer
- Branch plant key: integer
- Trans type key: integer
- Purchase month number: two digit integer
- Purchase day of month: two digit integer
- Purchase year number: four digit integer
- Currency Type: string with standard currency codes such as USD for US dollars
- Unit cost: real number (use Double data type in Access meta data)
- Quantity: integer

Data Source 2: Excel spreadsheet

- Customer vendor key: integer
- Item master key: integer
- Branch plant key: integer

- Purchase date: string with a format MM/dd/yyyy
- Trans type key: integer
- Currency Type: string with standard currency codes such as USD for US dollars
- Unit cost: real number (use Double data type in Excel meta data)
- Quantity: integer

2. ETL Operations

You should perform the following validations on each field and load validated records into the *Inventory_Fact* table of the inventory data warehouse.

- Reject a record if any field is null.
- Reject a record if any field value does not match its data type.
- Reject invalid dates: the combination of month, day, and year should be a valid date (including leap year processing) that exists in the *Date_Dim* table.
- Reject invalid foreign key references: the Customer vendor key, branch plant key, transtype key, and Item master key must be valid references to rows of the respective tables in the inventory data warehouse.

After validation, you should perform the following processing steps. These steps will enable the data to be loaded into the *Inventory_Fact* table of the inventory data warehouse.

- For data source 1, the month, day, and year fields should be used to find a matching row in the *Date_Dim* table in the inventory data warehouse. After finding the matching row, the *Date_Key* value in the *Date_Dim* row should be used for the *Date_Key* value in the *Inventory_Fact* table.
- For data source 2, the Purchase Date field should be parsed into its day, month, and year components. These components should be used to find a matching row in the *Date_Dim* table. Then, the *Date_Key* value in the matching *Date_Dim* row should be

used for the *Date_Key* value in the *Inventory_Fact* table. See the explanation in the following section about parsing dates in Excel data sources.

- The *UnitCost* column in the *Inventory_Fact* table is computed as the currency conversion factor times the Unit Cost field in the data source. The CREATE and INSERT documents contain a currency conversion table that you should use to retrieve the conversion rates in your Pentaho transform designs.
- The *ExtCost* column in the *Inventory_Fact* table is computed as the Unit Cost (after currency conversion) times the Quantity. In Pentaho, you need to use a Calculator step. In the calculation line with Quantity as a field, you must put Quantity as the B (right) field. Otherwise, Pentaho will convert the other field to integer before the calculation. According to the Pentaho documentation, "Calculator takes the data type of the left-hand size of the multiplication (A) as the driver for the calculation." If you put Quantity as the A field, the unit price after conversion will be converted to integer before the calculation giving the wrong result.

You should use Pentaho steps for the validation, processing, and loading. The data integration exercise provides background for most of the tasks in this assignment. To insert rows into the Inventory_Fact table, you should use a Table Output step.

3. Parsing Dates with Excel Data Sources

Parsing dates is more complex for dates in Excel data sources. In the data integration exercise, you were able to use a string data type in the Microsoft Access Input step and then use a select values step to perform the parsing. This approach will not work with dates in an Excel data source.

To begin, you should use Date as the data type in the Excel file (Microsoft Excel Input step). Do not use String as the data type as you used for the date fields in the Microsoft Access

Input step in the ETL exercise. You should make sure that the data type is set to Date by editing the schema of the Microsoft Excel Input step.

To parse dates in an Excel data source, you need to use a Select values step to convert the date field values into strings. In the Select values step, you need to alter the meta-data for dates field. Make sure that the type is string and the format parameter is exactly as shown ("MM/dd/yyyy"). The output date field in the Select values step can then be parsed by a Split Fields step like the parsing in the ETL exercise. Note that the regular expression must include the correct date component separator (/). In addition, the order of the new fields in the Split Fields step should match the parsing order (month, day, year). If you use a different order (for example, day, month, year), you will have an order error in the next step in the Merge Join. You will have rows rejected in the Merge Join because the parsing order does not match the output field order.

4. Merge Join Order

To assist with the assessment of your assignment, you should join the tables in the following order. Note that many other correct orders exist, but the assessment requires this order.

- Date_Dim
- Trans_Type_Dim
- Cust_Vendor_Dim
- Item_Master_Dim
- Branch_Plant_Dim
- Currency_Dim

5. Debugging Advice

It can be difficult to find some errors in a transformation as some of the error conditions indicate. If you experience errors that you cannot resolve, you should start a new transformation and then add/test steps one-at-a-time. You can use “preview this transformation” to see the results of the partial step to determine if the number of rows is correct. For multiple input steps like a merge join, you should add one input step at a time.

6. Grading

Two parts will assess your performance: a quiz and a self-evaluation. The quiz is designed to test your understanding of the assignment requirements. Since some quiz questions will involve execution of your transformations, you should be ready to incrementally execute

your transformations when taking the quiz. You need to execute your transformations against a clean version of the Inventory Data Warehouse tables. The self-evaluation requires snapshots for transformation and step metrics. You will receive 50% if you provide screen snapshots for each transformation. The other 50% will be determined by your response to the assignment quiz.

7. Submission

You should take screen snapshots of the complete transformation design pane, the execution of the final transformation showing the number of rows inserted into the Inventory_Fact table, and the step metrics showing the results of each step. You should take one set of screen snapshots for each transformation. Also, you should complete the module 5 assignment quiz.