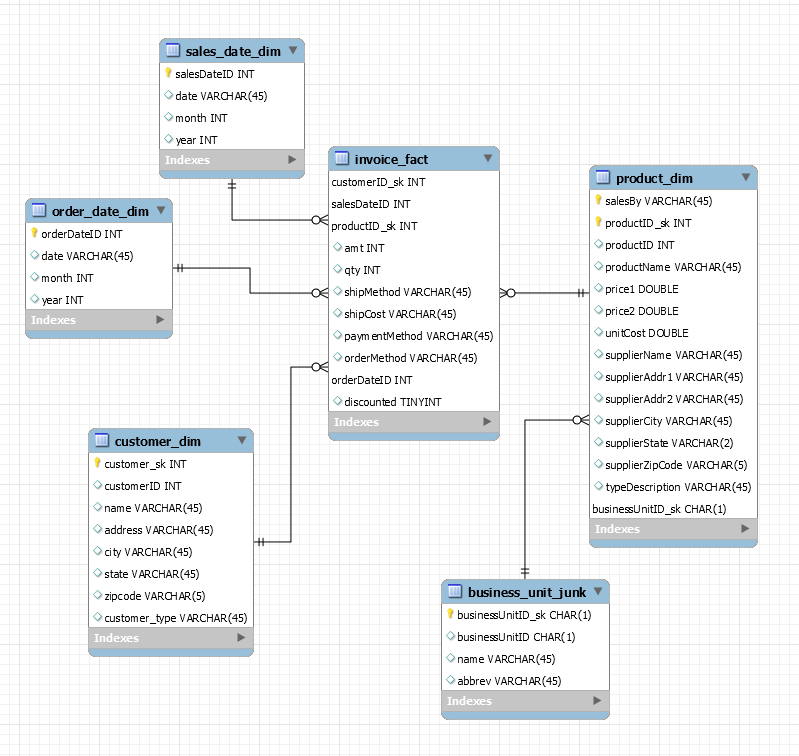
### ISTE Data Warehousing Lab 3

Ellie Parobek

Ryan Garcia

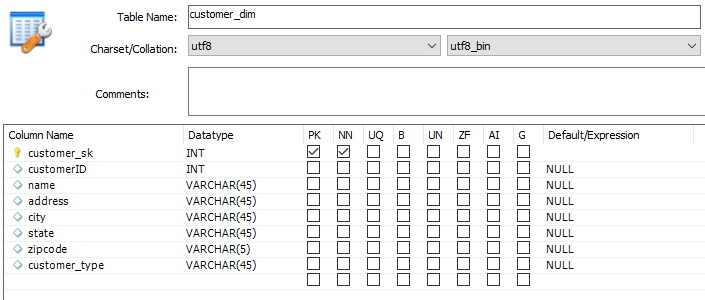
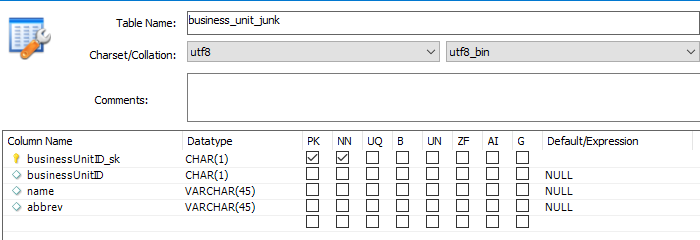
John Warner

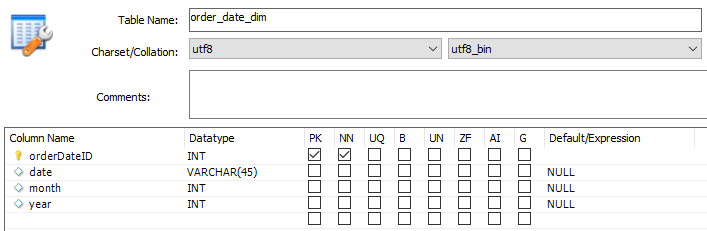
### Star Schema

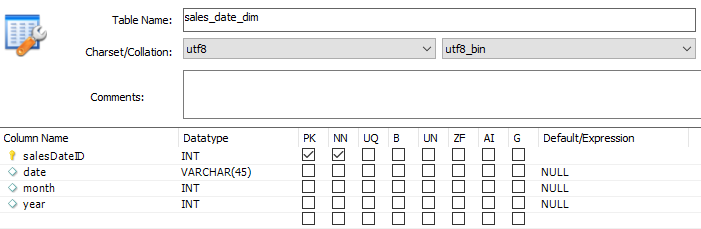


### 

### Final Tables





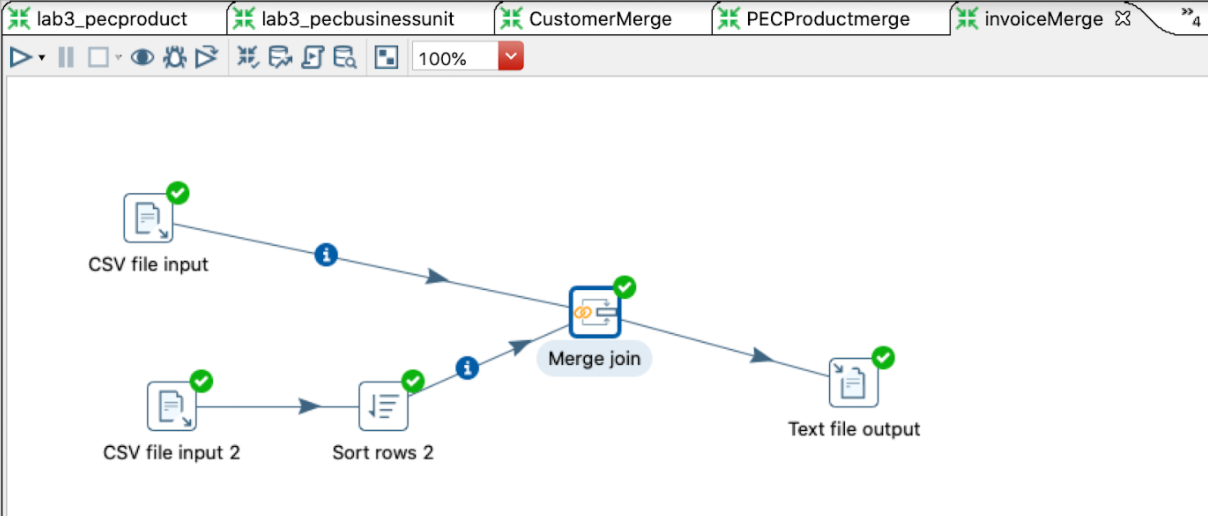


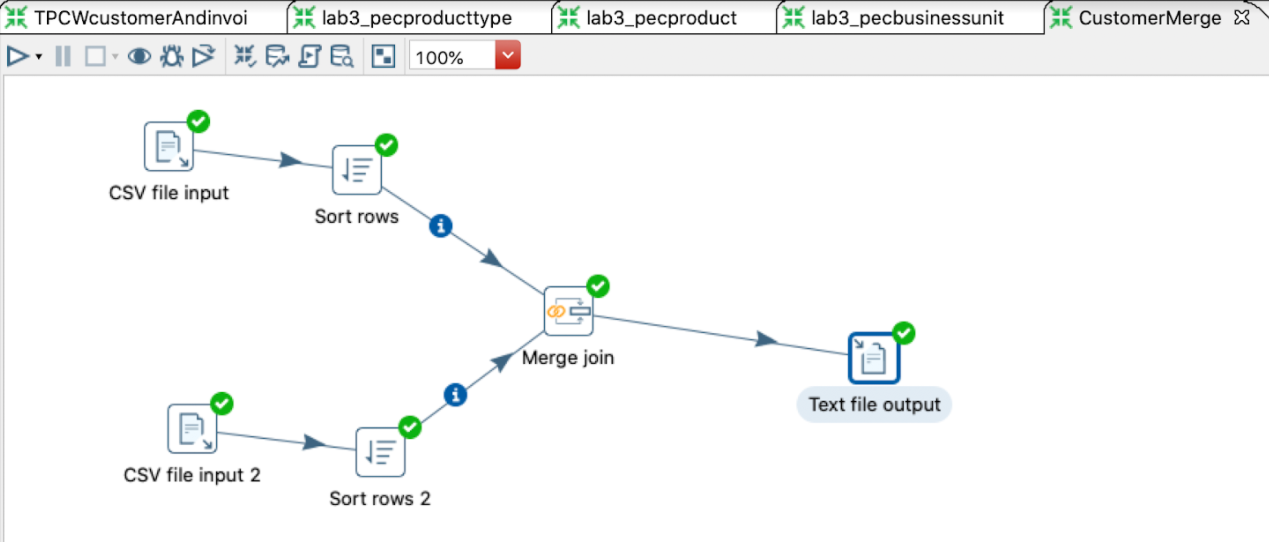
### 

### 

### 

### Transformations Screenshots

InvoiceMerge.ktr

CustomerMerge.ktr

### 

### 

### 

ProductMerge.ktr

### 

PECcustomer.ktr

### 

TPCWProduct.ktr

### 

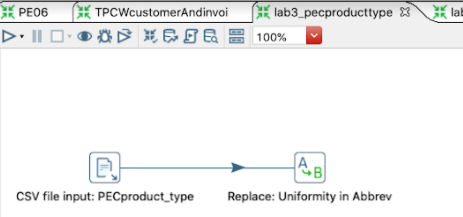
PECBusinessUnit.ktr

### 

PECproduct.ktr

### 

PECproducttype.ktr



### 

TPCWcustomerAndinvoice.ktr

### 

### Data Cleansing

Clean files are located under the ‘Cleaned Files’ directory. All ktr files are located under the ‘Pentaho Files’ directory.

|  |  |  |  |
| --- | --- | --- | --- |
| **File (.csv & others if applicable)** | **Attribute** | **Problem** | **Resolution Strategy (attach code)** |
| TPCWcustomer | address | Commas and periods | TPCWcustomerAndinvoice.ktr |
| TPCWcustomer | address | Department #s and suite #s included | TPCWcustomerAndinvoice.ktr |
| TPCWcustomer | zip | Zip code with 4 digits | TPCWcustomerAndinvoice.ktr |
| TPCWinvoice | All attributes | Columns shifted | TPCWcustomerAndinvoice.ktr |
| TPCWinvoice | invoiceID | Record with only invoiceID | TPCWcustomerAndinvoice.ktr |
| TPCWinvoice | custID | Negative custID | TPCWcustomerAndinvoice.ktr |
| TPCWinvoice | date | Date in 2 diff formats | TPCWcustomerAndinvoice.ktr |
| TPCWinvoice | amt | Sales amt incorrectly calculated | TPCWcustomerAndinvoice.ktr |
| TPCWinvoice | productID | None existent product ID (399) | Manually removed |
| PECinvoice | salesDate | Incorrect date format | PECinvoice.ktr |
| PECinvoice | orderDate | Incorrect date format | PECinvoice.ktr |
| PECinvoice | shipMethod | Incorrect shipping method | PECinvoice.ktr |
| PECinvoice | amt | Sales amounts are incorrect | PECinvoice.ktr |
| PECinvoice | All Attributes | 1 record has column values shifted | Manually corrected using Notepad++ |
| PECproduct | unitCost | Units costs are not properly calculated | PECproduct.ktr &  manually running equation for each product |
| TPCWproduct | SupplierCityAndState | Needs to be split into two attributes | TPCWproduct.ktr |
| TPCWproduct | All attributes | Duplicate entry | TPCWproduct.ktr |
| PECcustomer | address | Unnecessary commas | PECcustomer.ktr |

### Data Transformation

All ktr files are located under the ‘Pentaho Files’ directory.

|  |  |
| --- | --- |
| **DM Table** | **Image Creation Process (attach code)** |
| salesDate | Used Google Sheets formulas to convert salesDate into unique ids and create month/year columns.  Create month column: =LEFT(B1:B, 1) Create year column: =RIGHT(B1:B, 1) Convert date to unique id: = TEXT(C2,"mmddyy") |
| orderDate | Used Google Sheets formulas to convert salesDate into unique ids and create month/year columns.  Create month column: (=LEFT(B1:B, 1) Create year column: (=RIGHT(B1:B, 1) Convert date to unique id: (=ARRAYFORMULA(REGEXREPLACE(A1:A,"/",)) |
| Invoice (TPCW and TPCE) | Used Google Sheets formulas to convert salesDate into unique ids and create month/year columns.  Create month column: (=LEFT(B1:B, 1) Create year column: (=RIGHT(B1:B, 1) Convert date to unique id: = TEXT(C2,"mmddyy")  Also added shipMethod, shipCost, paymentMethod, orderMethod, and orderDate to be consistent with the other invoice file (PECinvoice).  Also updated customerIDs to match the customer table. |
| Invoice (PEC) | Used Google Sheets formulas to convert salesDate into unique ids and create month/year columns.  Create month column: (=LEFT(B1:B, 1) Create year column: (=RIGHT(B1:B, 1) Convert date to unique id: = TEXT(C2,"mmddyy")  Also updated customerIDs to match the customer table. |
| All customer files | Manually merged together to remove duplicates into ‘customer.csv’ |
| Customer | Merged together to form TPCEcustomer which was merged with the other customers to remove duplicates into ‘customer.csv’ |
| Product | Merged together PECproduct and PECproductType with Productmerge.ktr |
| Product | Merged together TPCEproduct and TPCEproductType with Productmerge.ktr |
| Product | Merged together TPCWproduct and TPCWproductType with Productmerge.ktr |
| Invoice (TCPE) | Merged together TPCEinvoice and invoice with InvoiceMerge.ktr |
| PECProductType | Removed commas, abbreviations, and cleaned up with PECproducttype.ktr |
| PECBusinessUnit | Removed commas, abbreviations, and cleaned up with PECbusinessunit.ktr |

### Table Population

All tables populated using table data import wizard and loading in each designated csv file in the ‘Files Loaded’ directory.

|  |  |
| --- | --- |
| **DM Table** | **Table Population Process (attach code)** |
| Business Unit Junk | businessUnit.csv |
| Sales Date Dim | salesDate.csv |
| Order Date Dim | orderDate.csv |
| Customer Dim | customer.csv |
| Product Dim | PECproduct.csv, TPCWproduct.csv, TPCEproduct.csv |
| Invoice Fact | PECinvoice.csv, TPCWinvoice.csv, TPCEinvoice.csv |

### 

### Part 5 Sample Queries:

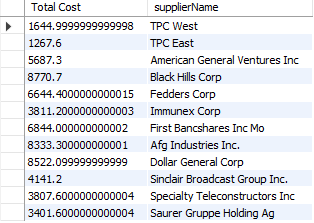
Show the total cost of products for each supplier

SELECT sum(unitCost) as 'Total Cost', supplierName

FROM product\_dim

WHERE supplierName <> ''

Group By supplierName;



The most frequent method of ordering a product from PEC

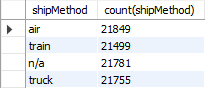
SELECT shipMethod, count(shipMethod)

FROM invoice\_fact

INNER JOIN product\_dim ON product\_dim.productID=invoice\_fact.productID\_sk

WHERE product\_dim.salesBy LIKE 'PEC' AND shipMethod <> ''

GROUP BY shipMethod;



Show sales from one division to another

SELECT SUM(invoice\_fact.amt) AS 'Total Sales', product\_dim.salesBy

FROM invoice\_fact

INNER JOIN product\_dim ON invoice\_fact.productID\_sk = product\_dim.productID

GROUP BY product\_dim.salesBy;  
 