Lab 3 – Extract / Transform / Load (ETL) Process

Overview

Your next step in the *TPC* data mart implementation is to cleanse and load data into a data mart implementation of your dimensional modeling star schema. You have requested various data extracts from the company's operational systems. The owners of the data have approved your requests and you will receive them. The extracts are in the formats that were convenient for the operational support staff to obtain. Therefore, you will need to ensure that you have the data that you need. Then, you will ETL these extracts to get the data into the format that you need for your data mart.

After completing this lab exercise you should be able to:

- Discuss the steps involved in a simple ETL process.
- Explain how a data cleansing tool can be used to review data and identify problems.
- Cleanse data.
- Reformat and load data into a data mart schema.

To do this lab you will need the following:

1) Your copy of the *TPC* case study, business rules, and ERD.
2) Access to a computer on which to do your data mart implementation.
3) Access to the databases you developed in Labs 1 and 2 for use by your team of developers.

Deliverables:

1) Dropbox - Answers to the questions in this specification in a separate document (MS Word).
2) Dropbox - An ER diagram showing your final physical Star Schema, developed in MySQL Workbench and pasted into a MS Word document.
3) Dropbox - A screenshot of your final tables and views from MySQL Workbench.
4) Dropbox – Screenshots of transformation and job diagrams from the Pentaho system.
5) Dropbox – the .ktr and .kjb files from your Pentaho work.
6) Dropbox – the .sas and .sas7bdat files from your SAS work
7) Dropbox - Your final MySQL Workbench model .mwb file
8) Dropbox – The SQL used to generate your database.
9) Dropbox – Cleansed PEC and TPCW csv files
10) Dropbox - A working MySQL database. Create a .sql file that contains the dumping of the SalesOrders_TeamNumber_2201 database.

Create a zip file: **Lab03_ TeamNumber_2201.zip** containing all files in the above 1) thru 10) to MyCourses Drop Box.

Part #1. ***Finalize Data Mart Dimensional Model***

Agree on Model

Get together with your development team members, and review and critique the data mart models your team members implemented in Lab #2. You may also want to look at the data being provided in part 2 of this lab.

Make sure that your database supports UTF-8 coding.

Be sure to document the rationale for all of your model decisions and your final schema design. Your team will defend these decisions at the project defense that will occur as the "final exam."

Part #2. ***Data Preparation – Data Analysis & Cleansing***

Step 2-1: Extract Data & Formats

There are three sources of data, one for each of the divisions of The Product Company.

TPC-E

The best source of TPC-E load data for the data warehouse is the OLTP database that you have created in Lab #1. You can extract this data directly from that database. This is a good time to consider a staging table, or another alternative is to use a view that looks like the staging table. The choice is yours.

Other Divisions

The other divisions are providing feeds of data that will have to be audited and reformatted to be loaded into the data warehouse.

Remember from the earlier specifications, they share the same customer and product hierarchies, but not the same customer or product IDs. Be sure to read the instructions carefully in the "Description of TPC" and Labs 1 and 2. Note especially that the customer data may overlap and that invoice numbers are not unique among locations. Some way will have to be considered to uniquely identify the invoice numbers. Product numbers are also not shared, but other product information should be mostly the same.

1. TPC-W

TPC-W will send six files. Study the files to consider any anomalies.

Customer data:

CustomerID

Name

Address

City

State

Zipcode

Customer type

40 Rows including the header

Product Data:

ProductID

Product Name

Price1

Price2

Unit Cost

Supplier Name

Supplier Address

Supplier city and State

Supplier zipcode

Product type ID

106 Rows with no header row

Invoice data:

InvoiceID

Customer ID

ProductID

Sales Date

Amount

Quantity

Discounted

89,548 rows including header row

Customer Type:

Customer Type ID

Type Name

5 rows including header row

Product Type:

Product Type ID

Type Description

Business Unit ID

15 rows including header row

Business Unit:

Business Unit ID

Name

Business Unit abbreviation

5 rows including header row

2. PEC

PEC will send seven files. Study the files to consider any anomalies.

Customer data:

CustomerID

Name

Address

City

State

Zipcode

Customer type

41 rows including the header

Invoice data:

InvoiceID

Customer ID

Sale date

ProductID

Amount

Quantity

Shipping method

Shipping cost

Payment method

Order method

Order date

Discounted

88,912 rows including header row

Product Data:

ProductID

Product Description

Price1

Price2

Unit Cost

Supplier Name

Product type ID

44 rows with header row

Customer Type:

Customer Type ID

Type Name

5 rows including header row

Product Type:

Product Type ID

Type Description

Business Unit ID

15 rows including header row

Business Unit:

Business Unit ID

Name

Business Unit abbreviation

5 rows including header row

On the product data feed, the unit cost and supplier name can be null. This indicates that the product is manufactured by PEC. The unit cost of a product can be calculated by using the manufacturing cost data in the following way. The manufacturing cost file has the cost of manufacturing a particular product in a particular month. By dividing that cost by the quantity sold of the same product in the same month, you can arrive at the unit cost of that product which can be updated to the data warehouse. Hint: the unit cost of a product should remain the same across all different months.

Manufacturing Cost:

Year

Month

Product ID

Manufacturing Cost

2311 rows including a header row

Review your data. You will need to extract it into usable file formats. You can use the Pentaho software to do this work.

With Pentaho, you should be able to find data that may be miscoded and duplicate that data into files. You may use a text editor to do any editing of this data.

Fill out **Table 1** in **Appendix A** to indicate the data problem(s) that you find. Explain how you will resolve the problem(s). Be sure to fully document all stages of your ETL activities. You will compile a report for your efforts as part of the final exam.

Step #2-2: *Agree on Model*

Again, now that you've had a chance to review the data, review and critique the data mart models that your team members developed. Agree on a model that you will implement. Document your rationale for your model choice. Your team will "defend" this choice at the project defense.

Part #3. *Data Preparation - Transformation*

Step #3-3.1: *Reformat the Data into "table images" for Loading*

Pentaho/SAS can be used for this or you may choose to use SQL to manipulate the data.

Fill out **Table 2** in **Appendix A** to fully document your activities since you will need to report on your efforts at your project defense.

Part #4. *Data Preparation - Load*

Step #3-4.1: *Load*

Load your cleansed and reformatted data into your data mart fact and dimension tables. You can use Pentaho, SQL or any other method that you find appropriate. Remember to handle your primary key, foreign key, and any other constraints appropriately.

Fully document your activities in **Table 3** of **Appendix A** since you will need to report on them at the project defense.

Part #5. ***Query Preparation***

Step #5.1: *Define User Queries*

Based upon your users' goals in Lab #2, select three report goals from the users and define a query for each. Define appropriate index(es), as necessary to support your query. Fully document your activities since you will need to demonstrate your query and its supporting indices at the final project defense.

Appendix A - ETL Data Staging Activities

1. Data Cleansing

File (.csv & others if applicable)	**Attribute**	**Problem**	**Resolution Strategy** (attach code)

2. Data Transformation

DM Table	**Image Creation Process** (attach code)

3. *Table Population*

DM Table	**Table Population Process** (attach code)