Data Warehouse Creation Tool

Steven Weigand
Computer Science – California Polytechnic State University
Spring 2008
Professor Dekhtyar

Table of Contents

Introduction

What is data warehousing?

A data warehouse is the actual data repository designed to facilitate reporting and analysis while data warehousing is the actual retrieval and interpretation of the components in the data warehouse. There are many reasons to use the data warehouse architecture. One benefit is the fact that the building of a data warehouse eliminates inconsistencies and redundancies from typical database schemas for easier analysis and interpretation of the data. Another benefit is that a data warehouse provides a common data model for all data of interest. Additionally, the construction of a data warehouse helps isolate the data of choice for a process known as data mining. Data mining is the process of sorting through large amounts of data and interpreting any relevant information. It is useful for making sense of data from large data sets that cannot be evaluated by hand.

Project Goal

The goal of the senior project was to create a tool that can allow a user to create and manipulate a data set into a data warehouse. The user creates the data sets by grouping the desired tables under a common name, separating them from the rest of the database. From here, the user would select his or her independent variables and dependent variables to create the data warehouse around. Afterwards, the user can view the status of the warehouse in three forms. The first way is an “Overview”, which explains the status of the warehouse via text. The second way is shows the sql code necessary to build the warehouse, and finally the third way represents the warehouse in diagram form.

What I Did

The following components make up each aspect of the data warehouse creation tool I have been working on for the past two quarters.

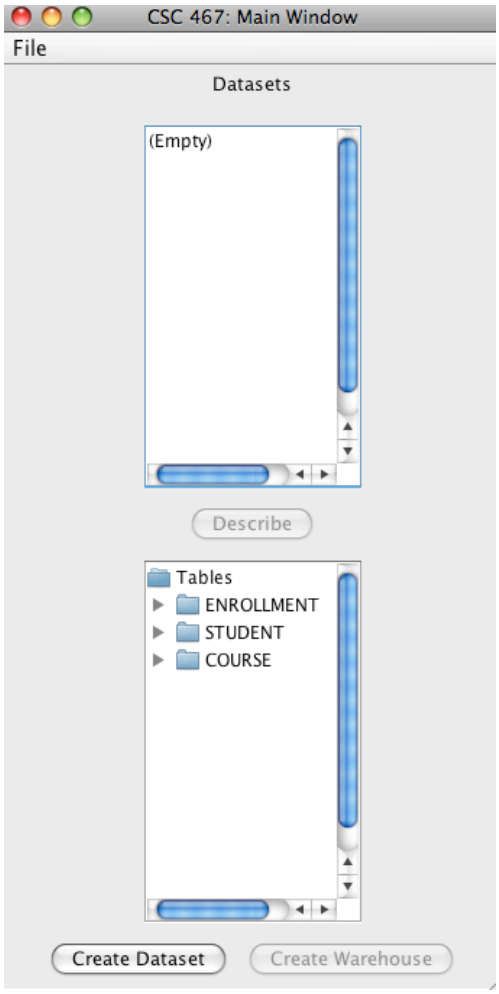
Main Screen

Purpose

The main screen of the tool is necessary to allow a user to be able to visually identify what tables the user has at his or her disposal. The tables will be displayed in tree form to make it easy for the user to also be able to view what columns exist within each table. Also, it allows a user to be able to visually identify what data sets have been created and what tables exist within each data set.

Description

Upon opening the tool, the user will be greeted with the "Main Screen" that contains too integral parts. The first list contains all of the data sets created by the user, while the second list contains all of the tables available to the user on his or her account. There are a number of options a user can utilize at this point. Under the "File" menu on the menu bar, a user can either login into his or her account or close the program. Logging into an account will fill the tables list with all available tables on the account and the data sets list will not be filled until the user creates some data sets from



the tables. Additionally, there are three buttons available for the user's disposal. The "Describe" button allows a user to view what tables exist within a given data set. The "Create Dataset" button will bring up a window allowing a user to organize tables into a data set. Finally, the "Create Warehouse" button will allow a user to begin manipulating a data set into a data warehouse.

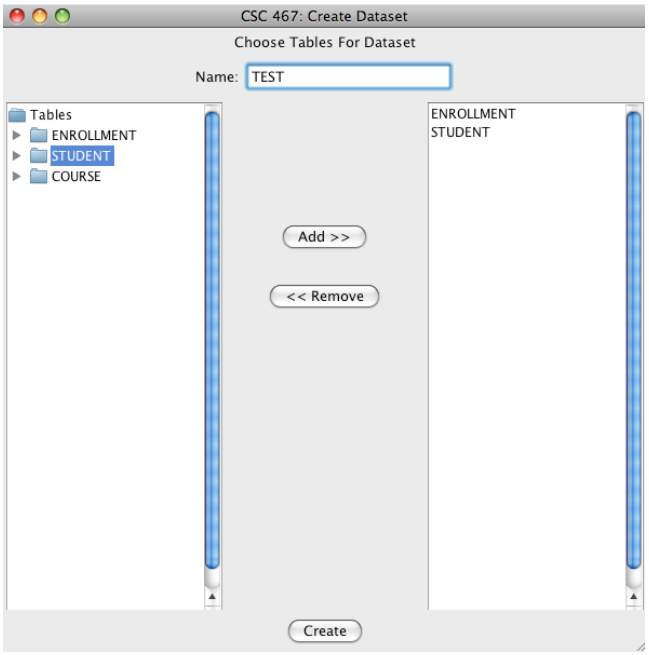
Creating A Data Set

Purpose

The "Create Set" window is important for allowing a user to be able to visually organize a data set by adding the desired tables to a list. Here the user view what tables have been added and also view what columns exist within each table to help the user decide, which tables belong in the given data set.

Description

When bringing up the data set creation window, the user is presented with an intuitive layout where the user can add tables from one list to another to create a data set. On one side of the window, the user can find all of the available tables on his or her account in a single list and an empty list on the other to which the user can add tables. In the center of the window, the user can find the two buttons necessary to perform the addition and removal of tables from the data set list. Finally, the top center of the screen contains the text



field in which a user can give the data set a name. Once the fields have been filled, the user can submit the data set and it will appear in the Main Window under the data sets list.

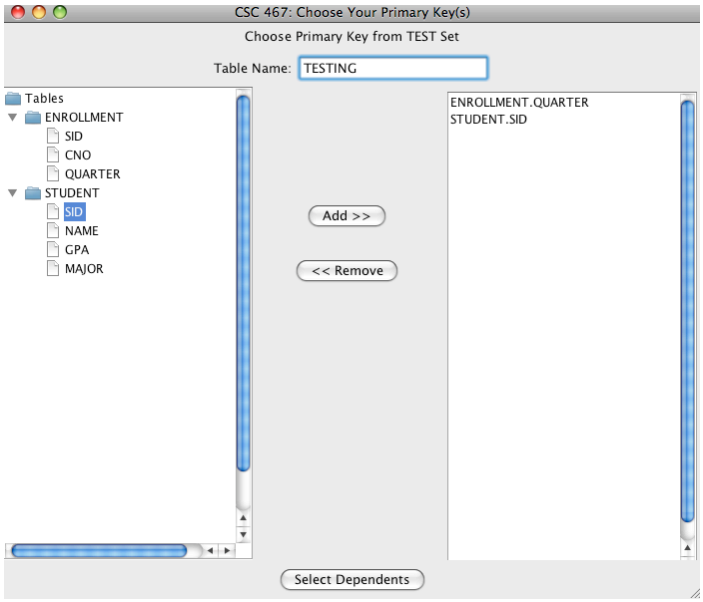
Selecting Independent Variables

Purpose

The select independent variables screen allows the user to select what columns will represent the data warehouse's independent variables. These values will eventually be utilized as the primary key of the main table within the data warehouse when the sql is created. The columns the user chooses from here are limited to whatever tables exist in the chosen data set.

Description

When bringing up the select independents window, the user is presented with a layout similar to that of the create a data set window where the user can add columns from one list to another to choose the independent variables. On one side of the window the user can find all of the available columns within the chosen data set in a single list and an empty list on the other in which the user can add columns to. In the center of the window the user can find the two buttons necessary to perform the addition and removal of columns from the independent variable list.



Finally, the top center of the screen contains the text field in which a user can give the main table of the data warehouse a name. Once the fields have been filled, the user can move on to the next window where he or she can select the dependent variables.

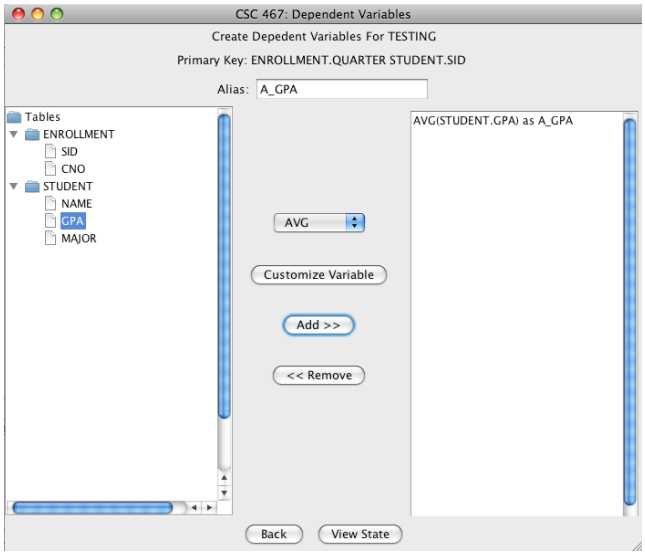
Selecting Dependent Variables

Purpose

The select dependent variables screen allows a user to choose columns to be numerically evaluated based on the independent variables. This allows the user to utilize some sort of sql aggregate function to aid in data analysis, such as SUM(SALES) if the user wants know the total sales based on whatever independent variables were chosen.

Description

When bringing up the select dependents window, the user is presented with a layout similar to that of the create a data set window where the user can add columns from one list to another to choose the dependent variables. On one side of the window the user can find all of the available columns within the chosen data set in a single list, minus whatever columns were chosen as independent variables, and an empty list on the other in which the user can add columns to. In the center of the window the user can find the two buttons necessary to perform the addition and removal of columns from the independent variable list. Additionally, the center of the window contains a drop down menu where a user can select an aggregate



function for the dependent variable. If the aggregate functions do not suffice, the user can choose to create a custom aggregate function using an expression of his or her choice. Finally, the top center of the screen contains the text field in which a user can give the variable an alias since the expression will not suffice for a column name. Once the fields have been filled, the user can move on to the next window where he or she can view the state of the data warehouse.

Viewing The Three Warehouse Panes

Purpose

The three-pane view allows the user to the view the warehouse in different forms before deciding to execute the sql code to construct it. This is important to allow the user a chance to ensure that the warehouse is in its proper form before committing it to their account.

Description

The three-pane view consists of the “Overview”, the “Code”, and the “Diagram”. Each give the user a different perspective of how the warehouse is put together in order to make a conscious decision on whether or not the design is correct before building it. The following describes what each pane has to offer.

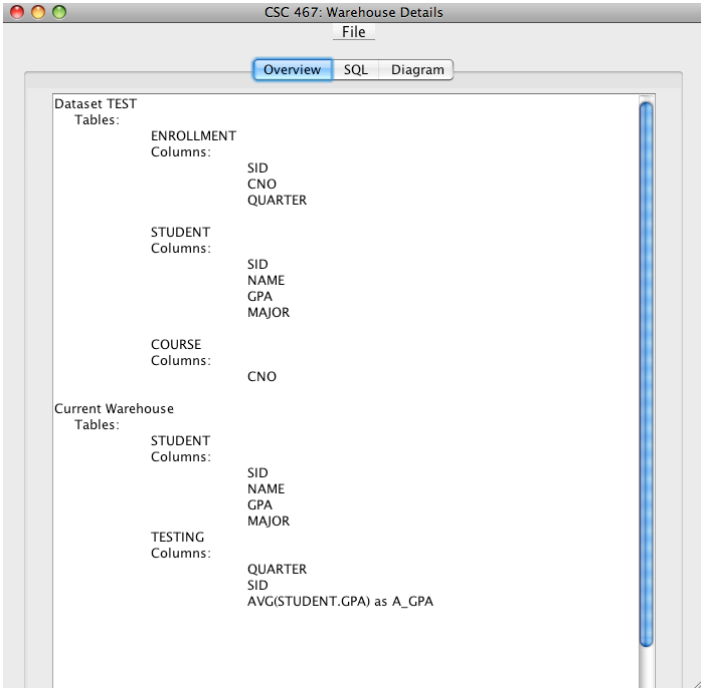
The “Overview” pane consists of only text that describes the warehouse from the data set to the tables making up the warehouse. This allows the user to view the original data set it was constructed from along with the layout of the warehouse. Here the user can reference tables and columns from the set that were not used and evaluate if any key components are missing.

The “Code” pane consists of only text as well, but contains the sql code required to create the data warehouse the user designed instead of comments. Here the user can see if code correctly represents the ideas the user had in mind before committing.

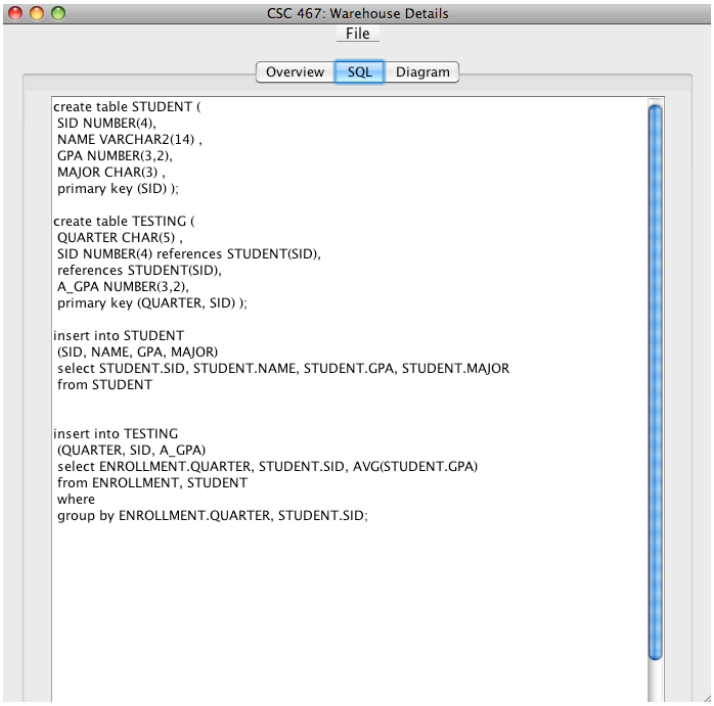
The “Diagram” pane visually shows a layout of tables and connecting lines show the user can see the relationships he or she created. This is important for the user to be able to visually see how the tables are laid out and see if any relationships were misinterpreted.

In addition to these panes, a “File” menu exists above the three panes giving the user four different actions to take. The user can “Save” the current state of what he or she as done, go “Back” to previous windows to reselect any independent or dependent variables, “Export SQL” to a file”, or “Execute SQL”, which would commit the warehouse to the database.

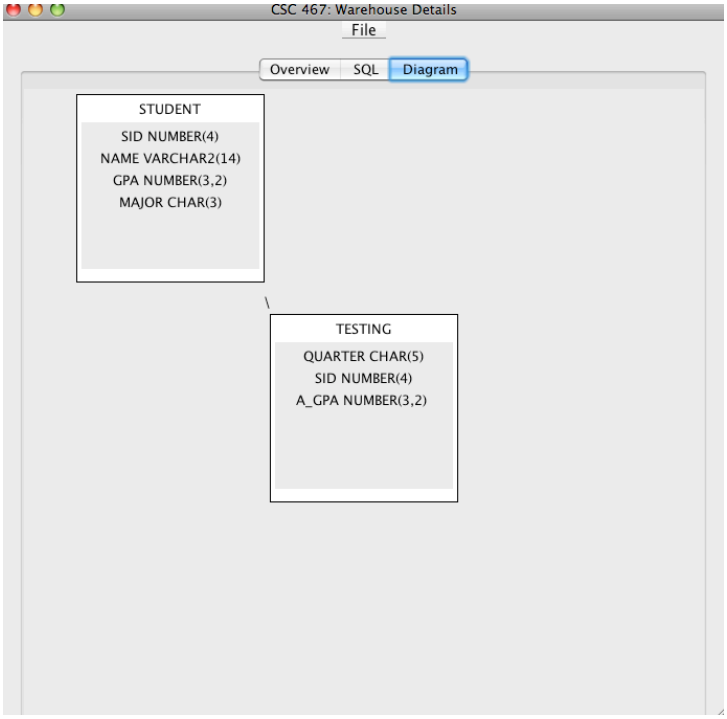
- Overview



- Code



- Diagram



Conclusion

My Overall Experience

As I completed this project, I feel that I have learned so much regarding the area of data warehousing. Before this senior project, I had sufficient knowledge regarding databases and database design from CPE 365 and CPE 366, but no training in the area of data warehouses. I had to spend my first few weeks of winter quarter researching the topic in order to have the understanding in order to be able to take on the project. Now that I have completed the tool, I feel a great sense of accomplishment that I have been able to become as knowledgeable as I have on the topic. Finally, I would like to thank Professor Dekhtyar for giving me the idea for this project and for being my advisor over the past few quarters. I appreciate all the help and enjoyed my time developing this tool.

References

Kumar, Vipin, Steinbach, Michael, and Tan, Pang-Ning. Introduction To Data Mining. Boston: Pearson Education, 2006.

Wikipedia. "Data Mining". 8 June 2008. Wikimedia Foundation, Inc. 10 June 2008 < http://en.wikipedia.org/wiki/Data_mining>

Wikipedia. "Data Warehouse". 9 June 2008. Wikimedia Foundation, Inc. 10 June 2008 < http://en.wikipedia.org/wiki/Data_warehouse>