CS 345 - Topics in Data Warehousing

Autumn Quarter, 2004

Assignment #2 – Due Tuesday, November 2

In this programming assignment, you will do extraction, transformation, and load for a simple data warehouse. The data warehouse stores performance statistics about major league baseball players. You will go through the ETL process twice: the first time you will load data from July, and the second time you will add data from August. This simulates real-world warehouses, which must be refreshed periodically over time. During the August load, you will encounter some changes to the players data source. Some new players will have been added, and the value of the `Weight` attribute for some of the existing players will have changed. You should handle this by using the "Type 2" slowly changing dimension strategy (creating additional rows in the Player dimension table for the changed players in order to preserve history).

This document describes the assignment at a fairly high level. More detailed information is available on the Assignments section of the course web page, including information about file formats for the provided data sources, how to connect to Oracle, hints and tips, test queries to verify your ETL worked properly, etc. **Additional information is available from http://cs345.stanford.edu/ in the "Assignments" section.**

Computing Resources Some support files for this assignment will be provided to make your task easier. They are located in the directory /usr/class/cs345/HW2, which can be accessed from any of the Sweet Hall UNIX machines (e.g. the elaines).

Each enrolled student has been provided with an account on Stanford's Oracle server. Information about logging in and interacting with Oracle is provided on the course web page, linked from the Assignments section. Your data warehouse should be constructed in Stanford's Oracle server; however, if you prefer to work with a different database system, you are welcome to do all the data staging on another system and then copy the final warehouse tables into Oracle once you are finished.

Schema The process being modeled in the data warehouse is batting performance by major league baseball players. There is one fact table (Batting) with the grain of one row per player per game. The dimensions of the Batting fact are Date, Player, Team, Opponent, and GameId. GameId is a degenerate dimension; it holds the unique game identifier assigned by the source system. The Team and Opponent dimensions are two "dimension roles" that reference the same dimension table (Team). Thus, there are three dimension tables in all (Date, Player, and Team).

The SQL statements to create the fact and dimension tables and load data into the Date dimension are provided for you. The following files can be found in the HW2 directory:

- **createTables.sql** Creates the fact and dimension tables.
- **dropTables.sql** Drops the fact and dimension tables.
- **insertDates.sql** Inserts rows into the DateDim dimension table.

An easy way to execute these scripts is to log into sqlplus and use the @ command, like this:
`@/usr/class/cs345/HW2/createTables`
The dropTables script is provided in case you make a mistake during your extraction process and want to start over from scratch.

Data Sources You will extract data from three different data sources.

1. **Team File:** Information about the major league baseball teams is provided in a comma-delimited file named `teams.csv` located in the HW2 directory. The comma-separated file has four fields: City, Team Name, League, and Division. These fields correspond exactly to the attributes of the Team dimension.
2. **Player Database:** Information about the major league baseball players is provided in two Oracle database tables, `babcock.players_v1` and `babcock.players_v2`. The `v1` table represents the state of the source system at the time of the initial extraction, and the `v2` table represents the state of the source system at the time of the follow-on extraction. Thus you should use the `v1` table when loading the July data and the `v2` table when loading the August data.
3. **Web Site:** The batting fact information comes from the ESPN web site. A wrapper program that crawls the web site and extracts the necessary data is provided in the HW2 directory. The wrapper is written in Perl, and the file name is `extract.pl`. To execute the wrapper script from one of Stanford's UNIX systems, use the following command syntax: "`perl /usr/class/cs345/HW2/extract.pl 1`". The "1" tells the wrapper to extract July data; replace the "1" by a "2" to extract August data.

What To Submit Submissions for this assignment will be performed electronically. You should create a directory in your Stanford UNIX account to hold the files you create for this assignment. The directory should include (1) a README file describing the steps in your ETL process, and (2) any code that you write, such as C or Java source files, SQL scripts, sqlldr control files, etc. Since many of the necessary tasks (surrogate key assignment, data transformations, etc.) can be performed in various ways, describe in your README how you performed each task.

When you are ready to submit your assignment, "cd" to your assignment directory and then run:
`/usr/class/cs345/submit`
If you discover an error after submitting, you can re-submit by running the submit script with the `-replace` command-line argument.

Steps To Follow

1. Log in to Oracle and change your password.
2. Run the `createTables.sql` and `insertDates.sql` scripts.
3. Load the Team dimension from the `team.csv` file.
   - Assign surrogate dimension keys for each dimension row.
   - Get the data out of the file and into the database.
4. Load the Player dimension from the `babcock.players_v1` table.
   - Assign surrogate dimension keys for each dimension row.
   - Transform the `Height` and `Weight` attributes to the proper data type.
5. Run the `extract.pl 1` script and load the Batting fact with July data.
   - Reconcile differences in the choice of natural key for Team between the Web Site and Team File.
   - Determine the appropriate surrogate dimension table keys to use as foreign keys for each fact row by maintaining a mapping from natural keys to surrogate keys.
   - Construct the derived `Singles` and `TotalBases` fact columns for each fact row.
   - Get the data out of the files created by `extract.pl` and into the database.
6. Run the test queries to validate that your July extraction was successful.
7. Load the Player dimension from the `babcock.players_v2` table.
   - Don't insert new dimension rows for existing, unchanged players.
   - Detect the new and changed players and create new dimension rows for them.
8. Run the `extract.pl 2` script and load the Batting fact with August data.
9. Run the test queries to validate that your August extraction was successful.