SWEG 6508 Data Warehouse

Semester Project

This semester project will give SWEG 6508 students the experience of developing a working data warehouse using a commercial database management system and development tools. Students will be assigned to a group. Each group will choose a large dataset from <https://www.kaggle.com/datasets> or some similar website. The group will manually download the dataset from the website. This data will be extracted, transformed, and loaded into a data warehouse and analyzed to help address a problem or issue identified by the group.

# Stages of the Project

Groups should first submit project ideas for approval. This proposal should include:

1. A narrative description of the issue or problem the group hopes to address - at least 2 paragraphs. Include a list of Key Performance Indicators (KPI’s) your group will use to address the issue or problem. The professor can help to refine the project and proposal. Students should not continue working on the project unless the proposal has been approved by the instructor.
2. Groups should design the data warehouse schema by defining dimensions and fact tables. A diagram should be created to document this step. Students must use a specific modeling tool such as MS Visio or LucidChart.
3. Groups should then identify a target DBMS that will host the data warehouse. The target DBMS could be a DBMS such as Oracle, SQL Server, MySQL, Postgres, Amazon Redshift or other. Groups will also need to decide on a hosting environment for this target database. Options include using the Oracle Cloud, Google Cloud, Amazon Web Services or Microsoft Azure, or hosting on a local computer or laptop though not recommended.
4. Groups should then download the Kaggle dataset which they previously chose and create ETL transformations for extracting and loading this data into the data warehouse schema. Tools such as Pentaho Data Integration (Kettle), Alteryx, Talend, Microsoft SQL Server Integration Services, etc. can be used. To document this work, take a screen picture of each major transformation module and add a short note describing what is happening in this stage of the ETL process. Once the ETL Process has been specified, groups should populate a data warehouse schema with data.
5. Groups should then create a basic “Dashboard” application that displays and incorporates at least 3 different data representations such as graphs, maps, heatmaps, charts, tabular/crosstab reports, etc.
6. This may be implemented in any analytics package such as Microsoft Power BI, Microsoft Excel, Pentaho CDE, Tableau, etc.

# Project Milestones

**March 23rd**: Project proposal due

**March 30th**: KPI’s and data sources identified. Dimensional modeling started and draft sent to professor.

**April 6th**: Final dimensional model sent to professor. ETL programming started.

**April 20th**: ETL Programming completed. Dashboard programming started.

**May 4th**: Dashboard programming completed.

**May 7th**: Presentation and final project report due by 6:30 pm.

For each milestone include a peer-review and self-review document:

# Final Project Deliverables

Each group should produce a document (MS Word) that includes the following:

1. A separate cover page listing the course, group member’s names and project title.
2. An introduction section, like the proposal section, including the narrative description of the business or organization used for the data warehouse being created and description of the source data. Be sure to include appropriate citations of each data source (web site, curator, date accessed, etc.)
3. The dimensional model diagram.
4. A description and screen pictures of each of the ETL processes.
5. Screen shots and brief descriptions of the final schema that the business analytics tools are working with.
6. Screen shots and descriptions of the analytics (at least 3) on the dashboard application developed based on the data warehouse data.
7. Descriptions of the tools (databases, analytics, ETL, programming languages, etc.) used to complete the project.
8. A narrative conclusion section that describes:
   1. the software and database tools the group used to coordinate and manage the project as well as carry out the programming tasks.
   2. the group’s experience with the project (which steps were the most difficult? Which were the easiest? what did you learn that you did not imagine you would have? if you had to do it all over again, what would you have done differently?)
   3. if the proposed benefits can be realized by the new system.
   4. any final comments and conclusions
9. A References list that provides the web sites and other sources for data, techniques, methods, software, etc. used to complete the project.
10. All the above materials should be arranged in **ONE** MS Word document to be submitted electronically.