Project Description

The course project on *Data Warehousing and OLAP* is a venue for students to practice building a simple data warehouse, and to subsequently design an OLAP application that accesses this warehouse. Students will again use the CBMS dataset from the **combined** provinces of Palawan and Marinduque to design a dimensionality model for their data warehouse; extract, transform and load the data from this dataset to their model; and formulate queries showing roll-up, drill-down, dice and slice operations. All decisions involved in designing the dimensionality model, ETL processes and query formulation, and the findings from evaluating the resulting data warehouse and OLAP application, will be documented in a technical report that students will be required to prepare and present as part of the output requirements.

Methodology

Students are to form teams with **3 - 4 members, subject to certain grouping constraints as discussed by your teacher**.

To proceed with this project, each team should conduct the following:

Step 1.   Dimensionality Modeling

Design a dimensional model containing **at least one (1) fact table** and **three (3) dimension tables** and following any of the multi-dimensional schema, namely *star*, *snowflake* or *constellation*. The schema can be implemented in MySQL or any of the more advanced DBMS (e.g., Microsoft SQL Server).

Criteria for grading the Schema:

- Correctness and Completeness.

- Relevance. The model should be able to support an organization, e.g., a local government or an NGO, in formulating appropriate intervention programs by providing data needed for planning. Conducting research into the different government programs for specific sectors (i.e., farmers, fishermen, OFWs, out-of-school-youth, women and children, elders) to deliver quality health care services, education support, alternative sources of livelihood, disaster response, and basic needs (e.g., electricity, water, food & income) is a necessary step in doing this project.

- Complexity. Evidence of clear appreciation and understanding of the benefits of a correct data warehouse schema in supporting OLAP analysis.

Step 2.   ETL Process

Find an existing tool to extract, transform and load the CBMS operational database content to the data warehouse. Teams may also opt to write their own simple ETL tool to perform this task.

Step 3.   OLAP Queries

Write a simple application in your chosen language, e.g., Java, C#, Visual Basic, or Python, that will allow the user to perform the following OLAP operations on the data warehouse: roll-up, drill-down, dice and slice. Appropriate design of a graphical user interface for the proper display of query results that will aid the user's analytical task is a must.

Step 4.   Evaluation

Test your dimensional model, ETL processes and queries. Testing should be efficient and effective.

Step 5. Prepare the Final Report

Using all the data from Steps 1 - 4, write your Technical Report. The outline is provided below.

1. Introduction. Give a brief description of the purpose of conducting this study. Describe the OLAP application that you built, its intended usage, and target users and beneficiaries.

2. Dimensional Model. Present and discuss your model.
   2.1. What are the contents of your fact table?
   2.2. What are your dimensions? Justify your choice of dimensions and facts.
   2.3. Which schema did you use – star, snowflake, or constellation? Why? What issues did you encounter in your model and schema design? How did you address these issues?

3. ETL Process. Discuss the processes of extraction, transformation and loading.
   3.1. If you used an existing tool, give a short overview of the tool. Why did you choose this tool?
   3.2. If you developed your own ETL tool, describe your code.
   3.3. For extraction, describe your data sources and their relevance to the model.
   3.4. For transformation, present the rules or functions you applied on the extracted data. What made you decide to apply these rules and functions? What issues did you encounter during transformation? How did you address these?
   3.5. For loading, discuss any additional constraints in the database schema. What issues did you encounter during loading? How did you address these?

4. OLAP Queries. Describe the main purpose of your application. What decision-making or analytical task(s) is the application intended for?
   4.1. Give a sample query for each of the following operations - roll-up, drill down, dice and slice.
   4.2. Describe each of the queries. Which dimensions are utilized (rolled up, drilled down, or constrained)?
   4.3. Provide sample output as appropriate or relevant to the discussion.

5. Results and Analysis. Discuss your test process, test data and results.

6. Conclusion. What are your learnings from conducting this project?

7. References. Reviewing literature on the domain of the data (government programs for the community), and technology (dimensionality modeling, star schema, ETL processes and tools, and OLAP operations) should be conducted to help you in doing your project and writing your paper.

Criteria for grading the Technical Report (A rubric will be used for grading.):
- Correctness and Appropriateness of the Dimensional Model
- Correctness and Completeness of the ETL Process
- Quality of the OLAP queries
- Efficiency and Effectiveness of the methodology (Test Process)
- Clarity of the discussion of results and analysis
- Evidence of Critical Thinking presenting the learnings derived from the conduct of the project (Conclusion), such as understanding the impact of the dimensional model and the ETL process to the types of OLAP queries that can be sufficiently supported
- Overall document presentation, e.g., format (title page, page numbers, sections, tables and figures), references, and language (spelling, choice of words and grammar)

<u>Final Deliverables</u>

The following final deliverables are required to be submitted:

1. Database Schema for the Data Warehouse
2. Source Code for the ETL process and the OLAP application
3. Technical Report following ACM publication format (www.acm.org/sigs/publications/pubform.doc)

- Softcopy of the deliverables must be submitted at <u>**12 midnight of March 20**</u> (MW classes) <u>**and 21**</u> (TH <u>classes).</u> Follow the file naming convention: **ADVANDB_<section>_<lastnames of members>.<ext>**

- Printed copies of the deliverables (if required) must be submitted during the first 15 minutes of the class on **March 21** (MW classes) and **March 22** (TH classes).

- Late submissions will receive 10 points deduction per day and 0 points for the presentation. Since March 23 has been announced as "No Classes" for DLSU, no late submissions will be accepted after March 22, 2016.

- Prepare a Powerpoint presentation to discuss your dimensional model, software application and OLAP queries, test methodology and results, and conclusion. The actual schedule for class presentation will be provided by your respective teacher.

- **Plagiarized works will automatically be given a grade of 0.0 for the course.**