CA: DT265 ADVANCED DATABASES

DR. BIANCA SCHOEN-PHELAN

28-10-2016

DUE DATE: week 13: Thursday, 08.12.2016

This assignment is worth 30% of your overall course mark.

The Task

Based on a collection of data of *your choice* downloaded from the **dublinked web page** construct an **ELT processing pipeline** that:

- Imports several CSV files into database tables (at least 2)
- Combines the data into a meaningful **relational** dataset
- From the **relational** representation of the data generate a **dimensional** model (star schema) of the data
- Your ETL pipeline is expected to make use of procedures, functions, triggers and rely on dynamically built SQL.

You should produce a **representational reports** (see template) from the dimensional data warehouse that highlights the **unique insight** generated by your combination of your data using a dimensional query or a data cube.

Clearly show your progression through the different staging areas of your ETL pipeline(in the code and the report)!

Objectives

The objectives of this assignment are to use SQL programming techniques to create an ETL processing pipeline that demonstrates:

- You understand the requirements and issues of the ELT process
- Your ability to use programmatic features of the RDBMS engine
- Your ability to create normalised and dimensional data models

Deliverables

The following deliverables are expected:

1. A database backup of all components of your processing pipeline and raw data (i.e. the SQL file)
2. A document explaining the structure of **your processing pipeline**, highlighting the **key code** and explaining **how to run it ("README file" section in the report)**. The document should also highlight the **key SQL programming constructs** used in your code. Use a **diagram** to illustrate your processing pipeline.
3. An explanation of the **insights** you obtained from your querying of the star schema.

Do not use the same data sets that you are using for your HDip project. This assignment focuses on the ELT/ETL processing pipeline and your modelling abilities.

Getting Started Hints and Tips

- Upload from CSV to tables
- Create a table from a select statement
- Check all values in a column against allowed values
- Find the best method of representing each stage schemas or separate databases

The Assignment will be evaluated on the basis of:

#	Category	Weight	In order to do well here I have to….
1	Initial investigation into data set to be used	10	Provide a detailed explanation of the datasets and the information provided in each. Detailed analysis of the requirements for an ELT pipeline.
2	ERD Design	10	Design a complex ER design (approx. 10 entities) that has been well described.
3	Dimensional Design	20	Dimensional design fully matches the previously described ER design and is semantically sound for the purpose of analytical querying (see 6).
4	Implementation	30	Implementation has been fully described and all necessary files have been provided. No issues replicating the student's work. README fully supports implementation.
5	Goal achievement	25	An automated ELT pipeline has been fully realised using procedural SQL wherever possible.
6	Insights	5	Novel insights due to the unique mix of datasets and dimensional query/or cube have been provided.

Resources

dublinked: http://dublinked.ie/