CSCI 5408, Winter 2018

Assignment 4: DW/OLAP & ETL Exercises Using IBM Cloud

(Issue: Feb 27, Due: Mar 15, 11:59PM)

- TA: Trishla Shah (trishla@dal.ca), Marker: Akash Rupareliya (akash@dal.ca)
- Ass4 Tutorial: Mon Mar 05, 4:05-5:25PM, Room: Chemistry 226
- Ass4 Help Hours: (All in room: CS 233)
  - Mar 08, 2:00 - 3:30PM
  - Mar 12, 4:00 - 5:30AM
  - Mar 14, 2:00 - 3:30PM

---

1. Objectives:

1) Learn how to build a data preparation pipeline for DW application (**Task 1**).
2) Learn how to design & implement a DW schema (**Task 2**).
3) Learn how to generate OLAP reports for the application (**Task 3**).

(*You are allowed to work in a team of 2 persons.)

2. Software Tools:

1) For Task 1 (ETL pipeline), you are recommended to use IBM Data Refinery, or an alternative, such as Google Cloud Dataprep or Tableau.
2) For Task 2, you are recommended to use IBM Db2 Warehouse, or an alternative, such as Tableau.
3) For Task 3, you are recommended to use IBM Cognos, or an alternative, such as Tableau or Dundas.com.

3. Task 1: Data Preparation ETL Exercise:

1) Choose proper application data sources for the selected DW application (refer to Task 2 on application data selection).
2) Identify data preprocessing tasks and design an ETL pipeline.
3) Use an ETL tool to implement the pipeline for generating target dataset(s). Your ETL implementation should perform the following tasks:
   - All SQL tables created from the source datasets should be in either 3rd or BC normal form, with identifiable relationship with other tables.
   - A workflow should also allow you to move data from the SQL tables to target flat file(s) based on partitioning predicates (merge, conditional splits, derived columns and aggregations) for the analytical task.
4) A brief description on the steps of using the tool to generate your result.
5) Your summary comments on the data ETL process.

4. Task 2: DW Design & Implementation Exercise:

1) Application dataset selection:

a) You can either use your own datasets, or other data sources, which is suitable for this exercise, e.g.
   - http://msftdbprodsamples.codeplex.com/releases/view/105902
   - https://northwinddatabase.codeplex.com/
   - http://mlearn.ics.uci.edu/MLRepository.html (e.g. "Adult" dataset)

b) Write up an application scenario (i.e. describe the application and the requirements, etc.) based on the selected dataset(s).

c) **Note:** The selected dataset(s) must have certain attributers can be used to form concept hierarchies for some DW dimension(s).

2) Use Snowflake/Star Schema to design a data cube and then upload the target dataset(s) on DB2 Warehouse or other after completing the ETL Process. Present the schema in your report.

5. **Task 3: OLAP Operations on DW**:
   1) Use IBM Cognos, Tableau or any other tool to Formulate and conduct each of the following operations: roll-up, drill-down, slice, and dice.
   2) Choose a proper visualization tool to present each result (include screenshots of the results).

6. **Report Requirements: (**The required sections)
   1) Application & Datasets: Describe the application by presenting a business scenario, including the type of business, application query examples, any other requirements), and the instances of data source tables (provide references).
   2) ETL Process: Description of the data preprocessing tasks, the designed ETL workflow, SQL tables created by processing the data at each process stage. Describe the relationship between them. Present the data flow tasks involved along with supporting screenshots showing the process results.
   3) DW Design & Implementation: Briefly describe your DW design consideration. Provide the DW schema diagram and an instance fragment of each table. Elaborate how various Fact and dimension tables are related to the corresponding tables in the OLTP datasets
   4) OLAP Queries & Reports: Each query is an ad hoc business question (in English) about some analytical information on the subject, which is then translated into OLAP operations, i.e. by choosing dimensions/concepts/values for producing a report, i.e. represent the retrieved information in an analytical screen (table) format.
   5) Summary: Provide a summary and observation of your work. Also, give feedback of tools which are used to complete the tasks.

7. **Submit your Ass4 report electronically:**
   1) Along with the report, provide the dataset link for task 1 (dataset and source link).
   2) Include generated PDF report in task 3 (OLAP operations)
   3) If any scripts or SQL statements must be run to set up the environment for task 1, task 2 and task 3, provide them along with their execution order and description.

4) Submit all files within one Zip folder on DAL Brightspace.

*** Plagiarism and Intellectual Honesty:** (http://plagiarism.dal.ca)
Dalhousie University defines "plagiarism as the presentation of the work of another author in such a way as to give one's reader reason to think it to be one's own." Plagiarism is considered a serious academic offense which may lead to loss of credit, suspension or expulsion from the University, or even the revocation of a degree.