Project

The Project contributes 40% to the total assessments of this unit, and requires two submissions to cssubmit during the semester.

Mid-project submission is an individual effort to receive feedback on your data warehouse design, due on Friday 23:59 pm 24th April (cssubmit). The deadline was extended from the originally scheduled deadline on 27th March, in order to allow students to have more time to complete the mid-project submission. It is worth 15% of the total assessments.

Final project submission can be an individual or a paired effort (i.e. to complete the final submission in a group of 1 or 2 people), due on Friday 11:59 pm 29th May (cssubmit). It is worth 25% of the total assessments.

Marking scheme will be available here as the semester progresses.

The overall objectives of this project are to build a data warehouse from real-world datasets, and to carry out basic data mining activities including association rule mining, classification and clustering.

Datasets and Problem Domain

For the datasets and problem domain, you have one of the following three options:

- **Prescribed datasets:** For this project, we would like to recommend the US Adult Income dataset (i.e. adult-training.csv) as the source of data for our data warehouse.
  - A local copy of the necessary files (adult-training.csv.zip)
- **Domain-specific datasets:** Kaggle.com host a multitude of data across different domains, governments such the USA and Australia all have made a tremendous amount of data available to the public. If you are motivated by a specific domain of interests, you can search these data repositories and discuss the datasets with lecturers and your lab facilitator.
- **You own datasets:** Datasets that you have collected or you have access to can be accepted as well. Talk to Zeyi and/or your lab facilitator about the suitabilities.

The project comprises of two sub-components, one on data warehouse design and implementation, the other on using the data for pattern discovery and predictive analytics.

1. Data Warehousing

Following four steps below of dimensional modelling (i.e. Kimball's four steps), design a data warehouse for the dataset(s) you have chosen.

1. Identify the process being modelled.
2. Determine the grain at which facts can be stored.
3. Choose the dimensions
4. Identify the numeric measures for the facts.

To realise the four steps, we can start by drawing and refining a **StarNet** with the above four questions in mind.

1. Think about a few business questions that your data warehouse could help answer.
2. Draw a StarNet with the aim to identify the dimensions and concept hierarchies for each dimension. This should be based on the lowest level information you have access to.
3. Use the StarNet footprints to illustrate how the business queries can be answered with your design. Refine the StarNet if the desired queries cannot be answered, for example, by adding more dimensions or concept hierarchies.
4. Once the StarNet diagram is completed, draw it using software such as Microsoft Visio (free to download under the Azure Education Link: https://aka.ms/devtoolsforteaching) or a drawing program of your own choice. Paste it onto a Power BI Dashboard.
5. Implement a star or snowflake schema using SQL Server Management Studio (SSMS). Paste the database ER diagram generated by SSMS onto Power BI Dashboard.
6. Load the data from the csv files to populate the tables. You may need to create separate data files for your dimension tables.
7. Use SQL Server Data Tools to build a multi-dimensional analysis service solution, with a cube designed to answer your business queries. Make sure the concept hierarchies match your StarNet design. Paste the cube diagram to your Power BI Dashboard.
8. Use Power BI to visualise the data returned from your business queries.

For mid-project submission, try to complete the above 8 steps. You can submit a Power BI Dashboard (.pbix file and its generated PDF file). Note you can add text descriptions to help illustrate the data warehousing process into Power BI Dashboard. Alternatively, you can write a document to include the business queries, the StarNet and query foot-prints, the Star/Snowflake Schema, the Cube Diagram, and the visualisation results to answer the business queries. Make sure you convert the file to pdf for submission.

Files to submit for mid-project submission:
The followings are the files needed for mid-project submission. **Please make sure that your files can be opened and run without needing a server connection.**

- The SQL Script file and the CSV files for building and populating your database.
- The solution project file (and its folder) of the SSDT analysis service multi-dimensional project.
- The Power BI file (.pbix) and a generated pdf file.
- A PDF file consists of the Design, Implementation and Usage of Data Warehouse.

All files need to be submitted to cssubmit.

Marking scheme:
[40 marks]
[5 marks] Concept Hierarchies and Corresponding StarNet
[5 marks] At least 5 business queries that the StarNet can answer

[5 marks] Star/SnowFlake Schema for DW design

[5 marks] Description of the ETL process for data transformation with code or screenshots
[5 marks] SQL Script file for building the database
[5 marks] SQL Script file for loading the datasets
[5 marks] Power BI visualisation corresponding to the 5 business queries
[5 marks] Coherence between the design and implementation, quality and complexity of the solution, reproducibility of the solution

****

Data warehousing exercises are often open-ended. In other words, there is almost always a better solution. You can interpret the scale of marks as:

5 - Exemplary (comprehensive solution demonstrating professional application of the knowledge taught in the class with initiative beyond just meeting the project requirement)
4 - Proficient (correct application of the taught concepts)
3 - Satisfactory (managed to meet most of the project requirement)
2 - Developing (some skills are demonstrated but need revision)
1 - Not yet Satisfactory (minimal effort)
0 - Not attempted.

****

2. Pattern Discovery and Predictive Analytics

One of the objectives of the data warehousing exercises in the first component is to produce clean, potentially aggregated, reduced or transformed data for pattern discovery and predicative analysis. In this second part of the project, we will demonstrate the typical data analytics processes using either **Weka** or other data analytic toolsets familiar to you (R or Python):

1. **Association rule mining**: select a subset of the attributes to mine interesting patterns. To rank the interestingness of the rules extracted, use support, confidence and lift. Explain the top 5 rules (according to lift or confidence) that have the "income" range/bracket on the right-hand-side; explain the meaning of the rules in plain English. Given the rules, what recommendation will you give to a person who wants to improve income (e.g. a person should receive more education to earn more)? If you use other data sets, explain the top 5 rules you obtain in plain English and also provide recommendation accordingly.
2. **Classification**: use the "income" range/bracket as the target variable or choose your own target variable, build two decision tree models: 1) one uses a list of attributes selected based on your understanding, 2) the other uses attribute selection based on information gain. Visualise the decision trees, and explain in plain English how the generated trees can be interpreted. Compare the performance of the two models using 10 fold cross-validation, and explain the evaluation results.
3. **Clustering**: run a clustering algorithm of your choice and explain how the results can be interpreted.
4. **Data reduction**: perform *numerosity reduction* using sampling and *feature reduction* with PCA or DWT. Train a model on the reduced data and train another model on the original data. Compare the performance of the two models using 10 fold cross-validation, and explain the evaluation results.

[Optional] You may use the test data set for model evaluation. You can also find the test data set in Kaggle or a local copy here

Marking scheme (Pattern Discovery and Predictive Analytics)
[30 marks]

[10 marks] Explain and interpret the top 5 association rules mined; based on the association rules, provide a recommendation for a person who wants to improve their in

[5 marks] Explain the attribute selection using Information Gain, and list the selected attributes.

[5 marks] Interpret and compare the results and performance of the Decision Tree Model with and without attribute selection
[5 marks] Interpret the clustering result (with respect to the target variable)
[5 marks] Explain the data reduction you have performed; compare the model trained on reduced data with the model trained on the original data.
[2 bonus marks] Put all three types of learning together (association rule mining, supervised learning and unsupervised learning), and interpret their relations in the cont

Files to submit for final project submission:

A report in PDF containing the four tasks (i.e. association rule mining, classification, clustering and data reduction) listed above. **If you work in a pair, only one submission is needed and the contribution of each team member should be clearly stated**. Clearly indicate the name and student number of yourself and the team member. The file needs to be submitted to cssubmit.