Project 1 - Data Warehouse Design

Project 1 contributes **25**% to the total assessments of this unit as an **individual** effort. The deadlines are recorded on cssubmit.

The overall objectives of this project are to build a data warehouse from real-world datasets and to carry out a basic data mining activity, in this case, association rule mining.

Datasets and Problem Domain

Prescribed datasets: the source data to design and populate the data warehouse in this project is based on the Olympic Dataset.

The Olympic Games represent the sole global, multi-disciplinary sports event, celebrated worldwide. Featuring participation from over 200 nations in more than 400 events spanning both the Summer and Winter Games, the Olympics serve as a platform for global competition, inspiration, and unity.

Paris *2024* will host the XXXIII *Olympic* Summer *Games*, 26 July to 11 August.

Paris 2024 - Official website
Paris 2024

Data sources

- olympic_hosts.csv and olympic_medals.csv originate from Olympic Summer & Winter Games, 1896-2022
- mental-illness.csv and life-expectancy.csv are sourced from Our World in Data ('DALYs' in the dataset stand for Disability-adjusted life years)
- Global Population.csv is obtained from the International Monetary Fund (IMF)
- Economic data.csv is derived from world-development-indicators
- list-of-countries_areas-by-continent-2024.csv is obtained from the World Population Review.

Note: All the datasets have been modified and differ from the original source. Please use the files provided below for your project.

Olympic.zip
archive

Special Requirements for CITS5504 and CITS3401

- **For CITS3401 students**, it is required that you identify at least ONE (1) client.
- **For CITS5504 students**, you must identify a minimum of TWO (2) clients. Examples provided below offer insights into potential scenarios
  - Clients may wish to query and analyze a common concept, but from different countries; for instance, Client A focuses on the USA and Client B on Australia.
  - Clients might be interested in querying and analyzing different concepts; for example, Client A could be exploring the relationships between the economy and the Olympic Games, while Client B is interested in understanding the connections between mental/physical health and the Olympic Games.
- **Both CITS5504 and CITS3401 students** must explain the reasons why the identified client(s) are important.

Data Warehousing Design and Implementation

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

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 to identify the dimensions and concept hierarchies for each dimension. This should be based on the lowest level of 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 Azure Education) or a drawing program/website of your own choice. A Hand-printed StarNet diagram is also welcome! Paste it onto an Atoti/Power BI Dashboard.
5. Implement a star or snowflake schema using SQL Server Management Studio (SSMS), or PostgreSQL, or other software. For the fact table and dimension tables, clearly state which ones are measures and dimensions, and indicate the dimension references.
6. Use Atoti 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.
7. Use Power BI/Atoti to visualise the data returned from your business queries.

Hint: You may need to convert the data type to help you generate meaningful queries.

Association Rule Mining

NOTE: you are not allowed to use Microsoft Visual Studio, R, or Weka to do association rule mining in this project. Python is the only option (for association rule mining) in this project.

Make sure you complete the relevant lab before attempting this task. The lab content may be helpful for you in completing this part.

Your objective is to assist a client in identifying significant patterns within the Olympic Games dataset. In order to demonstrate the application of association mining to this dataset, use this example to showcase the process and present the findings to the client.

Meanwhile, in the submitted PDF, you need to:

- Explain the top k rules that have suitable columns on the right-hand side based on a suitable metric, where k>=1.
- Explain the meaning of the k rules in plain English.

- Share insights derived from the mining results. If no meaningful rules are discovered, explore potential reasons for this outcome.
- Give the client at least THREE (3) suggestions on ***commerce*** based on the obtained results.

Answer the following question in your report

Some articles argue that

`"The data cube is an outdated technology."`

> For example, in a 2023 article, Albert Wong argued that "database cubes were popular in the early days of data warehousing, but they have largely been replaced by other technologies."

Do you agree or disagree with this point?

- If you disagree with this point, discuss your reason.
- If you agree with this point, discuss your reasons and explain at least one technology that can replace the data cube.

Requirements:

- Write a short **argumentative essay** to answer this question in your submitted PDF file.
- Use various forms of evidence, such as data, experience, facts, or literature to support your points.
- Word limitation: 400-500.

What to submit

1. A PDF report structured according to the above 7 steps to explain

- the fact table and dimension tables design and their ER diagram.
- the concept hierarchies for each dimension and how they can support the granularity of your data warehouse queries;
-

the Extraction Transformation and Loading (ETL) process to show how the source file can be loaded into the data warehouse;

- how multi-dimensional cubes are used to facilitate the roll-up and drill-down analysis;
- how PowerBI/Atoti can assist with query result visualisation.
- descriptions of the association rule mining process and results.
- schema visualisation of your data warehouse

2. All scripts you used (e.g. Python, SQL) with clear and well-structured comments.
3. Explanation and steps of the data cleaning/preprocessing/ETL process.
4. (If you use Power BI) The Power BI file (.pdf file).

All files need to be zipped up in a single zip file and submitted to cssubmit.

Marking scheme

[65 marks]

[5 marks] Schema of each Dimension and Concept Hierarchies for each Dimension

[5 marks] Corresponding StarNet to illustrate query capabilities of the DW

[5 marks] At least 5 types of business queries for each client that the StarNet can answer

[5 marks] Star/SnowFlake Schema (Fact Table) for DW design

[10 marks] Data cleaning/pre-processing/ETL process for data transformation with code or screenshots or explanation

[5 marks] Coding quality and structure

[5 marks] Visualise corresponding to the 5 business queries with appropriate tools

[5 marks] Association rule mining meaningful set-up

[5 marks] Interpretation of top rules and suggestions

[5 marks] Overall report quality

[5 marks] The quality of the answer to the given question

[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 with clear understandings demonstrated)

3 - Satisfactory (managed to meet most of the project requirements)

2 - Developing (some skills are demonstrated but need revision)

1 - Not yet Satisfactory (minimal effort)

0 - Not attempted.

****

Bonus Marks (15 marks)

Up to 15 marks of bonus marks can be awarded for (but not limited to)

- Outstanding ETL process.
- Query design. ***Queries are complex and meaningful.***
- Visualization. ***Visualise results using alternative methods or tools instead of Atoti. All selected chart types are appropriate. The design and layout are beautiful and suitable.***
- Association rules mining. ***Implementing the Association Rules Mining part without using Python packages, except `pandas` and `numpy`.***
- What-if analysis. ***Implementing meaningful and correct What-If analysis in the project. You may use Python or Excel to complete this bonus task.***

Can I use ChatGPT?

You are more than welcome to use ChatGPT, but please include the ChatGPT suggestions in your report and justify why you adopt or not adopt its suggestions. It is not compulsory to use ChatGPT. However, as a data science professional, it is strongly encouraged to learn to criticise and be conversant in modern new tools that can potentially enhance your productivity.

Note: Using ChatGPT without reference is unacceptable.

The following example is a BAD example that may risk receiving a 0 mark.

Using "the second person" in your report is weird.

```
server <- function(input, output) {

  # Your data preprocessing and plotting code here

  metrics_data <- data.frame(type="videoviews",
           values=data_preprocessing$video.views) %>%
  rbind(data.frame(type="videoviews_thelast30days",
                   values=data_preprocessing$video_views_for_the_last_30_days)) %>%
  rbind(data.frame(type="subscribers",
                   values=data_preprocessing$subscribers)) %>%
  rbind(data.frame(type="subscribers_thelast30days",
                   values=data_preprocessing$subscribers_for_last_30_days))

  output$metricPlot <- renderPlot({
    selected_metric <- input$metricSelect

    metric_plot <- ggplot(metrics_data, aes(x = type, y = values)) +
      geom_boxplot(data = filter(metrics_data, type == selected_metric),
                   notch = TRUE, outlier.colour = "blue", outlier.shape = 18, outlier.size = 4) +
      stat_boxplot(data = filter(metrics_data, type == selected_metric),
                   geom = "errorbar", width = 0.15) +
      scale_y_log10(
        breaks = trans_breaks("log10", function(x) 10^x),
        labels = trans_format("log10", math_format(10^.x))
      ) +
```

Next
Project 1 Q&A

Last updated 13 hours ago