ASSIGNMENT 12.1: Explore Data Warehouses

Start Assignment

- Due Tuesday by 11:59pm
- Points 100
- Submitting a file upload
- File Types pdf
- Available until Apr 11 at 8am

Format

Method: Individual.
Materials: **freeformatter.com; (https://freeformatter.com/xml-formatter.html)** and R Studio (Posit)

Overview

In this assignment, you will explore analytical databases, specifically those designed using the Kimball Star Schema design approach rather than the Inmon design approach.

Start by creating a new R Project and within the R Project, create a new R Notebook with the naming pattern *CS5200.LastNameF.DWA.Rmd*. Write all of your answers in markdown within the R Notebook and eventually submit a knitted PDF of the notebook.

> You may use any resources available on the web, including AI assistants such as ChatGPT or Bard, but you cannot directly copy from any sources and you must acknowledge the use of any resources. Any AI generated content must be edited and carefully checked to ensure it is correct and relevant. Naturally, you need to understand what is generated -- after all, the purpose of an assignment is to learn. We will check for plagiarism, so do not copy directly and without acknowledgement -- any direct copy is presumed to be intentional and constitutes an academic integrity violation with commensurate consequences as outlined in the syllabus.

Question 1 (30 Points)

Data warehouses are often constructed using relational databases. Explain the use of fact tables and star schemas to construct a data warehouse in a relational database. Also comment on whether a transactional database can and should be used to house an OLAP. Lastly, why would an organization

use a relational database with a star schema rather than a dedicated NoSQL database specifically engineered for data warehousing and analytics?

Question 2 (30 Points)

Explain the difference between a data warehouse, a data mart, and a data lake. Provide at least one example of their use from your experience or find out how they are generally used in practice. Find at least one video, article, or tutorial online that explains the differences and embed that into your notebook.

Question 3 (40 Points)

After the general explanation of fact tables and star schemas, design an appropriate fact table for Practicum I's bird strike database. Of course, there are many fact tables one could build, so pick some analytics problem and design a fact table for that. For example, you might want to build a fact table for "birdstrike facts", such as the average and total number of bird strikes by time periods (weeks, months, years), or bird strikes by region or airport and, again, broken down by time period. This might be useful for time series analysis or to determine seasonality for bird strikes which might then inform notices to pilots. Be sure to explain your approach and design reasons.

Just design it (ideally by creating an ERD for it); you do not need to actually implement it or populate it with data (of course, you may do so if you wish in preparation for the next practicum).

Submission

Submit a knitted PDF of your R Notebook.