Cairo University
Faculty of Computers and Artificial Intelligence
SIS314 Advanced Database –2025/2026 – Dr. Shaimaa Galal

Assignment 1: Data Warehouse

Students number: Max 5	**Deadline**: **12 Nov. 2025**

Consider the following Kaggle Datasets:

a. Financial dataset for fraud detection
URL : https://www.kaggle.com/datasets/ealaxi/paysim1
A synthetic dataset that contains transactions based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country. The original logs were provided by a multinational company, who is the provider of the mobile financial service which is currently running in more than 14 countries all around the world.

b. Anime recommendation dataset
URL : https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database
This data set contains information on user preference data from 73,516 users on 12,294 anime. Each user is able to add anime to their completed list and give it a rating and this data set is a compilation of those ratings.

c. EPL 21-22 **English Premier League**
URL : https://www.kaggle.com/datasets/azminetoushikwasi/epl-21-22-matches-players
This dataset contains all the stats of **English Premier League season 2021-22.**

d. IPL Indian **Premier League**
URL : https://www.kaggle.com/datasets/azminetoushikwasi/epl-21-22-matches-players
Dataset includes Data of all the IPL Seasons (637 matches)

1. Design a star schema for a database of your choice. We stated four datasets above, however, you are free to choose any other dataset other than the "Sales dataset" studied in lectures. Try to find challenges or questions to be asked when dealing with the chosen dataset. For example, if we consider the Premier league dataset, we might need an answer for the following questions:
   - **Discover the weak points** of any team.
   - **Suggest players need to be sold,** based on performance analysis.
   - Nominate **Player of the season**

2. Define dimensions, fact table(s) you will include in your star of snowflake schema. Minimum number of dimensions are 4 and number of measures are 2. Date dimension is a must to include in your schema.

3. Consider data are being provided every day to the system administrator in CSV file (the one from Kaggle dataset). Design SQL stored procedure or SQL statement to load your data from CSV file and sends an email for a predefined email (system administrator) with the loading process result (Success or Failure).

4. Design SQL Job to run your "SQL stored procedure or SQL statement" everyday at a predefined time ONLY to add the new data loaded to your file.

5. Implement SCD type 6 for at least two fields of any table, then write a stored procedure to read the source data using Incremental load.

Assignment Printable Deliverables:

a. Cover Page contains the following (**Group ID, DB Source Name, Group names, IDs, and Emails**)

b. Source ERD (interested tables) – You build your star schema.

c. Motivation for creating your star schema.

For example, we are creating the Product sales star schema (for Adventure works) to analyze the sales profit statistics for each product and its categories and sub-categories in specific intervals of time.

d. Star Schema Model (Dimensional model).

e. Schema Description (Dimensions, Dimension Levels, and Measures):

****For each dimension:**

Write its query and the query description and if there are levels/ hierarchies in the dimension state them.

****For Fact table:**

Write its query and the query description (also write the measures equations- if applicable)

Assignment Code Deliverables:

a. SQL stored procedure or SQL statement to load your data

b. Job that runs (a).

c. Print screen of the sent email.

d. Stored procedure implementing the SCD type 6.