TopGitHubRepos: Database Documentation

Tables:

T1: language
Source link: https://www.kaggle.com/datasets/sujaykapadnis/programming-languages
Licencing: Public Domain
Columns used: Language name, Language type, Language year.

T2: author
Source Link: https://www.kaggle.com/datasets/sujaykapadnis/programming-languages
Licencing: Public Domain
Columns Used: Author Name
Columns Generated: Birth Country
GPT Prompt: for each of the names in the list of software engineers I am about to give you, associate a country to each one. If no country is found, give a country of NULL

T3: git_repo
Source Link: https://www.kaggle.com/datasets/parulpandey/most-starred-github-repositories
Licencing: CC0 Public Domain
Columns Used: ID, Repo Name, stars, language Name, Last Commit

T4: country
Source Link: https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023
Licencing: Attribution 4.0 International (CC BY 4.0)
Columns Used: Country, GDP, Minimum wage, Population

T4: author_lang
Foreign key referencing lang_name in language table
Foreign key referencing author_name in author table

Business rules:

Each language can be developed by zero or more authors, and each author can develop one or more languages.
- The linker table between author and language enforces this many to many connection

Each git hub repo is programmed with either zero or one languages.
- The foreign key in git_repos links to language enforcing the many to one constraint (or can be null)

Each language is identified by a unique name.
- In the language table, the primary key is language name, enforcing this constraint.

Querries

1) This select statement shows all authors from countries starting with the letter C
2) This select statement counts how many git repos belong to a specific language type
3) This select statement counts how many git repos belong to a specific language type, it also displays their average year of development
4) This select statement shows the average number of employees fore countries
5) This select statement shows how authors developed language contributes to a number of git hub repos and the difference in stars from their lowest to highest. The dml statement rounds the star difference and rounds it to its nearest thousandth.

Stored procedures

1) This procedure is able to add a new language to the dataset
    a. Params
        i. IN new_lang VARCHAR(50) = the name of the new language to be added
        ii. IN new_type VARCHAR(50) = the type of the new language
        iii. IN new_lang_year VARCHAR(50) year = year of the year of development
        iv. IN new_author VARCHAR(50) = author name
        v. IN new_birth_country VARCHAR(50)) = birth country of new author

2) This procedure is used to update the GitHub table.
   It removes all repos that use a specific language
   It updates all repos of a cetertain language to rename that language. This cascades to other tables
    a. Params
        i. IN lang_to_remove VARCHAR(50) = name of language to remove
        ii. INOUT num_removed INT = returned num of removed rows
        iii. INOUT num_updated INT = returned num of updates rows
        iv. IN lang_to_update VARCHAR(50) = lang name to update
        v. IN new_lang_name VARCHAR(50) = new lang name