This report explains the process of creating a relational database in the Phase-I for the project. The dataset we selected is Meta Kaggle (available at: <https://www.kaggle.com/datasets/kaggle/meta-kaggle>). The report contains brief description about the dataset, a basic relational model, and details about the program that loads a subset of the dataset into the database.

Note: The program code is provided in the root folder. The program can be executed by following the README.md instructions. The questions from the root directory contain method calls from the other files in the “funcs” subfolder to make the code more manageable and readable. The global settings are provided in “globals.py” inside the “funcs” subfolder.

Task: Select one or more datasets to be used in the rest of the project. You may use existing datasets or may compile your own.

**Selected Kaggle Meta**

Task: The final dataset needs to be large (~50M tuples in a relational database), and interesting enough so you can perform meaningful queries and mine meaningful information from it. You need to provide a link to the dataset and a detailed description of the data.

**Give a link; and dataset description; and explain how it could be interesting. Can be run using analyze data code providing selected data files.**

Task: Give a meaningful relational model to faithfully represent the dataset

**Create a relational model; perhaps ERD?**

Task: Give a program to load the dataset. (Note that phase III of the project will involve data cleaning, so you do not need to be overly concerned with cleaning your data at this stage.)

**Write a program to load the dataset without cleaning; perhaps ignoring foreign key constraints.**

Lists Tables here to what can constitute meaningful and more than 50 million rows;

1. Users

Rubric Checklist: Personal, remove before submission:

|  |  |  |
| --- | --- | --- |
| Sr. | Rubric | Sufficient? |
| 1 | The meaning of the dataset as a whole is thoroughly described and each field or attribute is explained in detail. | No |
| 2 | The relational model is clearly described along with a clear explanation of how the data in the dataset maps to the provided model. | No |
| 3 | The code is well-commented and documented. It is clear how to configure and run the program. | No |
| 4 | The relational model uses appropriate tables, columns, data types, and primary/foreign keys that faithfully represent all elements of the dataset. | No |
| 5 | After configuration, the loading program handles the steps of creating the database and loading all the data without manual intervention. | No |

Note: Instructions to run the code are provided in the README.md file in the root folder.