**PROJECT 2**

**REPORT**

Jasmine Blakely, Jesus Paz, Sheena Knott

**Introduction**:

This study is an exercise in ‘Extract, Transform and Load’ (ETL), whereby data is extracted and cleaned from various sources, processed and reorganized for ease of usage, and finally stored as a database.

In this study, crowdfunding in entertainment industry related to audio-visuals and food, is considered. Each campaign is associated with a goal of raising a set amount of funds. While some campaigns are successful in reaching the goal or exceeding it, others are not, and some are cancelled.

Two Excel files contain information regarding various projects and points of contact. Information is parsed and processed from both. Four new files are created which can be linked to one another. A new database is then created to store the four files.

**Part 1:**

Excel file ‘crowdfunding’ is read into pandas DataFrame. Each unique name in fields,

‘category’ and ‘subcategory’, is given an identity code, *viz.*, ‘category\_id’ and ‘subcategory\_id’, respectively.

**Two** new files are created: one with information regarding ‘category’, and the other containing data regarding ‘subcategory’.

In the original file, ‘crowdfunding’, features not directly relevant to crowdfunding are eliminated. Data-types for attributes, ‘goal’ and ‘pledged’ are changed from ‘object’ to ‘float’. This **modified file** is named ‘campaign’.

All **three** files are converted into csv files.

Excel file, ‘contacts’ has each record in a dictionary, with keys and values in a dictionary as one attribute.

Data is parsed to create 4 columns: contact\_id, first\_name, last\_name, email.

This **fourth** file is subsequently stored as a csv file.

**Part 2:**

Data for storage in Database **‘cowdfunding\_db’**:

1. CSV file name: **campaign**

Number of records: 1000

Attributes: cf\_id, contact\_id, company\_name, description, goal, pledged, outcome, backers\_count, country, currency, launched\_data, end\_date, category\_id, subcategory\_id

1. CSV file name: **contacts**

Number of records: 1000

Attributes: contact\_id, first\_name, last\_name, email

1. CSV file name: **category**

Number of records: 9

Attributes: category\_id, category

1. CSV file name: **subcategory**

Number of records: 24

Attributes: subcategory\_id, subcategory

In PostgreSQL 12, using pgAdmin, a database called **crowdfunding\_db** is created.

Four tables are created, followed by importing data from each corresponding CSV file into each table.

To determine a primary key (attribute) for each table, a check for uniqueness is conducted.

Foreign keys are also determined to establish inter-table relationships.

Note: The primary and foreign keys are added **after** creating tables and importing data, to avoid encountering error due to potential erroneous choices of primary keys.

ERD

Information from the tables created in Part 1, including the primary and the foreign keys for each table was input into quickDBD, to create an Entity Relationship Diagram (ERD):

A screenshot of a computer

Description automatically generated

Table 1

Sample data in ‘**campaign**’

A screenshot of a computer

Description automatically generated

Table 2

Sample data in ‘**contacts**’

A screenshot of a computer

Description automatically generated

Table 3

Sample data in ‘**category**’

A screenshot of a computer

Description automatically generated

Table 4

Sample data in ‘**subcategory**’

A screenshot of a computer

Description automatically generated

**Conclusion**:

* Data entered into pandas DataFrame from two Excel files is parsed and processed. The objective is to assign one specific feature to each attribute in each file.
* The simplified data is then formed into four CSV files such that each has a relationship to at least one other file.
* The four files are stored as tables in a database ‘crowdfunding\_db’, so as to form a relational database.

The purpose of this operation is to enhance the understanding of data and to present the data in a simplified form desirable to the user.