CSCI-620: Introduction to Big Data - Project : Phase 1

ARPAN SHAH, Rochester Institute of Technology
DIPTANU SARKAR, Rochester Institute of Technology
LIPISHA NITIN CHAUDHARY, Rochester Institute of Technology
RITABAN BHATTACHARYA, Rochester Institute of Technology

This document presents the Phase 1 of the Project.

CCS Concepts: • **Information systems → Data management systems**;

Additional Key Words and Phrases:

ACM Reference Format:
Arpan Shah, Diptanu Sarkar, Lipisha Nitin Chaudhary, and Ritaban Bhattacharya. 2019. CSCI-620: Introduction to Big Data - Project : Phase 1. 1, 1 (February 2019), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This paper describes the dataset that has been selected for the phase 1 of the project and the relational model created for the schema designed for this dataset.

2 DATASET

The dataset has been chosen from Kaggle[3].

2.1 Dataset used

Beers, Breweries and Beer Reviews

2.2 Link to the Dataset

https://www.kaggle.com/ehallmar/beers-breweries-and-beer-reviews/version/2

2.3 Description of the Dataset

This dataset describes information about assorted beers, with details about breweries across the world, and the availability of these beers brands covering over the breweries. Additional to this the reviews from a myriad of customer drinking the beers is recorded and matched with the information about the beers and collated with respect to the region. The dataset is defined as a CSV (Comma Separated Values), which mainly comprises of three principal files namely: beers.csv, breweries.csv, reviews.csv.

Authors' addresses: Arpan Shah, as6999@rit.edu, Rochester Institute of Technology; Diptanu Sarkar, ds9297@rit.edu, Rochester Institute of Technology; Lipisha Nitin Chaudhary, lc2919@rit.edu, Rochester Institute of Technology; Ritaban Bhattacharya, rb5344@rit.edu, Rochester Institute of Technology.

XXXX-XXXX/2019/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.3.1 *beers.csv.*

- *Id :* an integer type which uniquely identifies individual types of beers.
- *Name :* a varying character type, which provides the name of the beer.
- *Brewery_Id :* an integer type, which uniquely identifies the brewery.
- *State :* a varying character type giving the state, in which the beer is available.
- *Country :* a varying character type giving the country of the state, in which the beer is available.
- *Style :* a varying character type, which supplies the information of the type of beer.
- *Availability :* a varying character type, which gives the availability of beers round the year.
- *Abv :* a decimal type, which gives the amount of alcohol present per volume in the beer.
- *Notes :* notes, a varying character type, provides additional information regarding the beer.
- *Retired :* a Boolean type, which gives true if the beer has been deprecated, else false.

2.3.2 *breweries.csv.*

- *Id :* an integer type which uniquely identifies an individual brewery establishment.
- *Name :* a varying character type, which gives the name of the brewery.
- *City :* a varying character type, which gives the city in which the brewery is present.
- *State :* a varying character type, which gives the state of the city, in which the brewery is present.
- *Country :* a varying character type, which gives the country, in which the brewery is established.
- *Notes :* a varying character type, which provides with additional supporting information about the brewery.
- *Types :* an array type, gives the type of brewery it is, with any subsidiary establishment present with it.

2.3.3 *reviews.csv.*

- *Beer_Id :* an integer type which uniquely identifies an individual beer whose reviews are to be presented.
- *Username :* a varying character type, which provides the username of the person who provided the review of a specific beer.
- *Date :* a date type variable, gives the exact date at which the review was provided
- *Text :* a text type variable, which describes the entire review of the particular username for the given beer.
- *Look :* a decimal type variable, gives the rating of how the looks of the beer is, on the scale of 5.
- *Smell :* a decimal type variable, gives the rating of how the smell of the beer is, on the scale of 5.
- *Taste :* a decimal type variable, gives the rating of how the taste of the beer is, on the scale of 5.
- *Feel :* a decimal type variable, gives the rating of how the feel of the beer is, on the scale of 5.
- *Overall :* a decimal type variable, which gives the overall rating given by the username for a particular beer.
- *Score :* a decimal type variable, which gives how the beer performed among the all it's users.

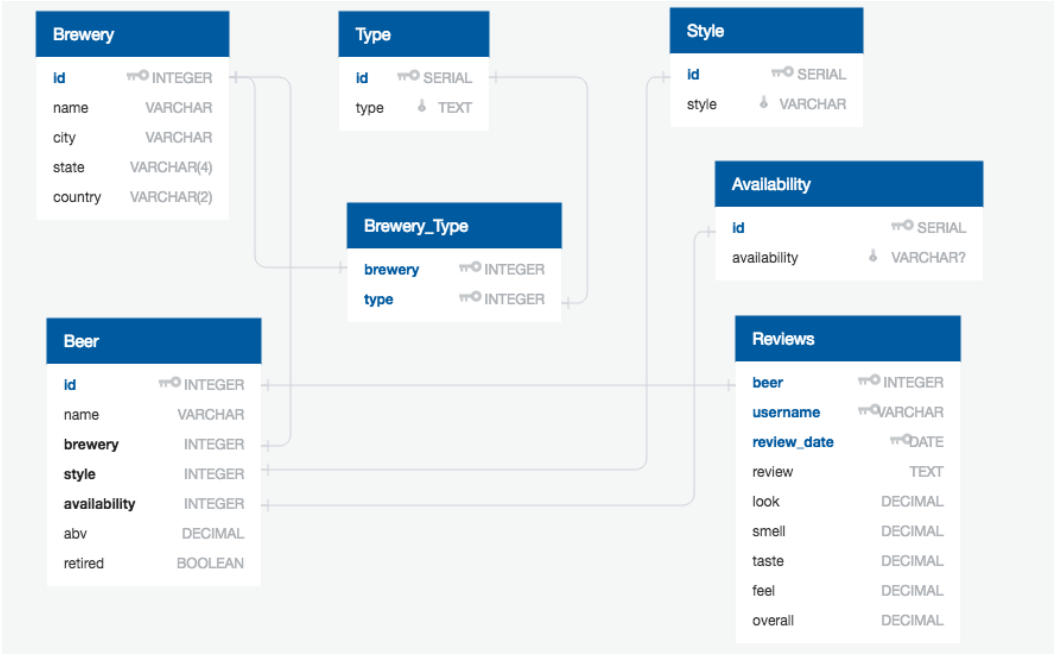

Fig. 1. Relational Model

3 RELATIONAL MODEL

The relational diagram was designed using Quick Database Design [6]. The sql scripts are provided separately in a .sql file along with this paper. The datasets were inserted into the database using Python[5] IDE, Pycharm[4]. For reading the datasets, the package csv[2] is used. For connecting to the plsql database, a package psycopg2[1] is used, Installation details can be found in the link http://initd.org/psycopg/download/.

3.1 Description of Relational Model

3.1.1 Brewery.

- *id :* a integer type which uniquely identifies the breweries independently. Created from Id attribute of breweries.
- *name :* a varying character type, which gives the name of the brewery. Created from Name attribute of breweries.
- *city :* a varying character type, which gives the city in which the brewery is present.. Created from City attribute of breweries.
- *state :* a varying character type, which gives the state of the city, in which the brewery is present. Created from State attribute of breweries.
- *country :* a varying character type, which gives the country, in which the brewery is established. Created from Country attribute of breweries.

The Notes attribute of breweries has been omiited from the design since it doesn't bring any significance to our design. The types attribute of breweries has been isolated to another table to properly categorize each brewery with the respective types.

3.1.2 Type.

- *id :* a integer type auto generated serial id for each type of brewery.
- *type :* a text type, gives the type of brewery it is, with any subsidiary establishment present with it. Created from distinct records of type attribute of breweries.

3.1.3 Brewery_Type.

- *brewery:* id attribute of brewery table.
- *type:* id attribute of type table.

The content of this table is based on which breweries can be categorized to which types.

3.1.4 Style.

- *id:* a integer type auto generated serial id for each style of beer.
- *style:* a varying character type, which supplies the information of the type of beer. Created from distinct records of style attribute of beers

3.1.5 Availability.

- *id:* a integer type auto generated serial id for each avalaibility season of beer.
- *style:* a varying character type, which gives the availability of beers round the year. Created from distinct records of availability attribute of beers

3.1.6 Beer.

- *id :* an integer type which uniquely identifies individual the beers. Created from Id attribute of beers.
- *name :* a varying character type, which provides the name of the beer. Created from Name attribute of beers
- *brewery :* id attribute of brewery table asscociated with this beer.
- *style :* id attribute of style table asscociated with this beer.
- *availability :* id attribute of availability table asscociated with this beer.
- *abv :* a decimal type, which gives the amount of alcohol present per beer. Created from Abv attribute of beers
- *retired :* a Boolean type, which gives true if the beer has been deprecated, else false. Created from Retired attribute of beers

The Notes attribute of beers has been omiited from the design since it doesn't bring any significance to our design. The State and Country attribute of beers has been omitted assuming that the beer will have the same state and country as its brewery. To properly category the style, availabilty and the brewery associated with it, the style and availability table has been isolated and the id's are used in this table, same for brewery.

3.1.7 Reviews.

- *beer :* id attribute of beer table associated with this review.
- *usrename :* a varying character type, which provides the username of the person who provided the review of a specific beer. Created from Username attribute of reviews.
- *review_date :* a date type variable, gives the exact date at which the review was provided. Created from Date attribute of reviews.
- *review :* a text type variable, which describes the entire review of the particular username for the given beer. Created from Text attribute of reviews.
- *look :* a decimal type variable, gives the rating of how the looks of the beer is, on the scale of 5. Created from Look attribute of reviews.

- *smell :* a decimal type variable, gives the rating of how the smell of the beer is, on the scale of 5. Created from Smell attribute of reviews.
- *taste :* a decimal type variable, gives the rating of how the taste of the beer is, on the scale of 5. Created from Taste attribute of reviews.
- *feel :* a decimal type variable, gives the rating of how the feel of the beer is, on the scale of 5. Created from Feel attribute of reviews.
- *overall :* a decimal type variable, which gives the overall rating given by the username for a particular beer. Created from Overall attribute of reviews.

The Score attribute of reviews has been omitted since it can be calculated with the other scores provided in the design.

REFERENCES

[1] 2018. PostgreSQL + Python | Psycopg. http://initd.org/psycopg/
[2] 2019. csv — CSV File Reading and Writing — Python 3.7.2 documentation. https://docs.python.org/3/library/csv.html
[3] 2019. Kaggle: Your Home for Data Science. https://www.kaggle.com/
[4] 2019. PyCharm: the Python IDE for Professional Developers by JetBrains. https://www.jetbrains.com/pycharm/
[5] 2019. Python 3.0 Release. https://www.python.org/download/releases/3.0/
[6] Dovetail Ltd. 2019. QuickDBD. https://app.quickdatabasediagrams.com/