Introduction to Big Data - Project (Phase-1)

Sharath Nagulapally, Harshraj Mahesh, Uddesh Karda, Sowrabh Nandamuri

February 2019

1 Introduction

Our aim is to select a data set and build a relational model for it based on the data available in the data set. We selected the Yelp data set which consists of attributes such as ratings, reviews, locations and other information about various establishments such as restaurants. The model is built is based on proper analysis of data and only needed data is chosen from the data set.

2 Files in the Data Set

2.1 business.json

This file consists the main information about the business. It contains attributes such as business ID which is an alphanumeric PRIMARY KEY and name, address, state, city, postal code which are of VARCHAR datatype. In addition, they contain attributes review count. Ratings are taken in a separate TABLE which has the business ID as FOREIGN KEY. Similarly, Address is also taken as another TABLE with values ID, state, city and postal code having business ID as FOREIGN KEY.

2.2 Business_category and category

The business.json file contains an attribute called category which is an array of values which tell what kind of business establishments it comes under. For this a new relation which category ID as PRIMARY KEY taken and category name for all available categories is made. The Business_category relation has business ID and category ID as FOREIGN KEYs which help to IDentify a specific business with a category.

2.3 User.json

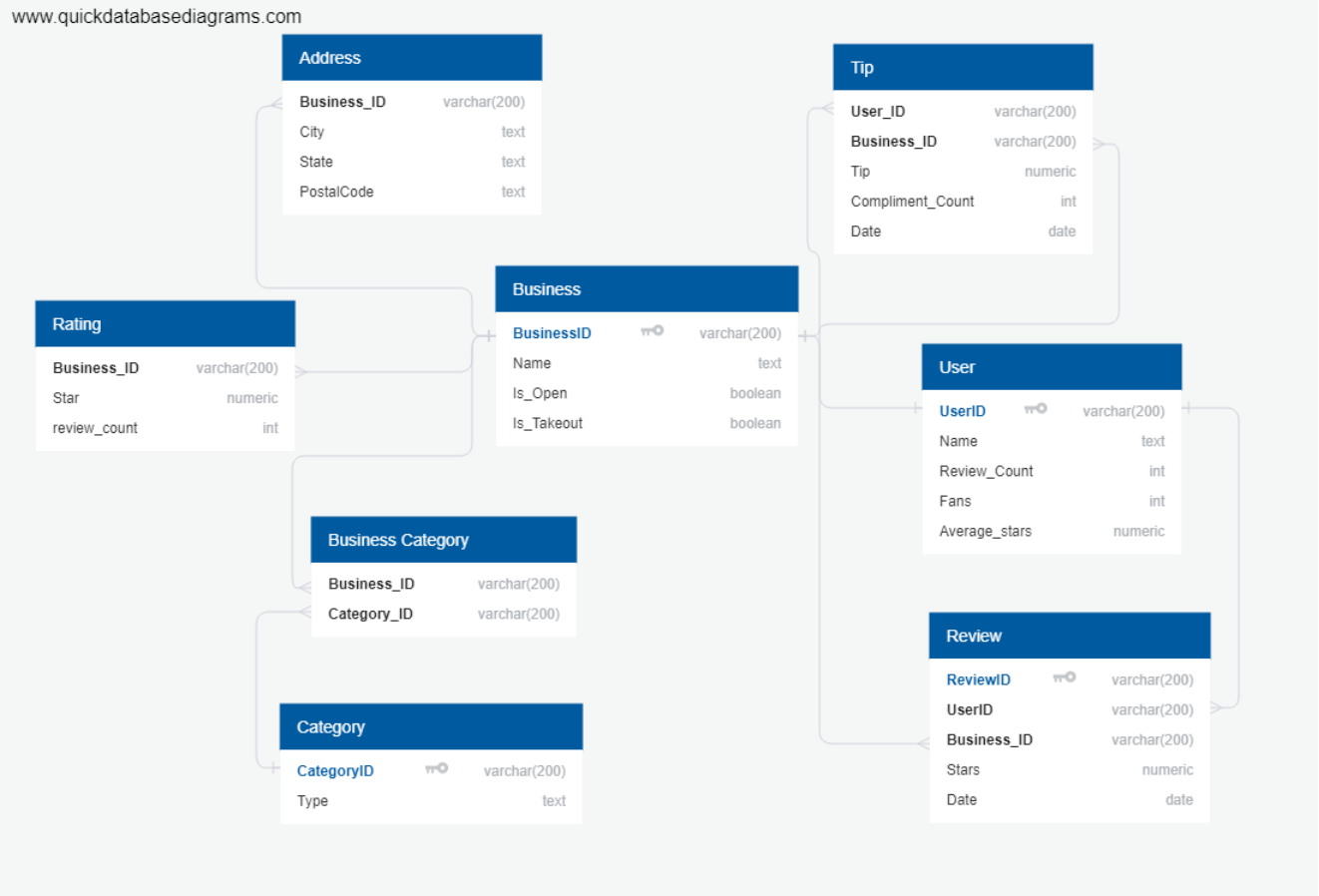
This file contains data about the users who have CREATEd an ID on the website. A specific user ID, name, review count, fans etc data is available. We are going to store the ID, name, review count, fans(shows how popular the user is) and average stars (shows how strict the user is in rating a business) in the User TABLE.

2.4 Tip.json

This file contains the tips given by each user to a specific business the compliment count and date of the specific tip given. The user ID and business ID acts as a FOREIGN KEY to IDentify a record uniquely. All these information is going to be stored in relation called Tip with the above attributes.

2.5 Review

This relation connects a specific user from the user relation and a business from the business relation as to what rating has a specific user has given for a specific business. It also contains the date on which this specific rating has been given.



This shows the relational model we designed from the given data from the yelp dataset.

3 Creating the Relational Model

Business
CREATE TABLE Business(BusinessID VARCHAR(200) PRIMARY KEY, Name TEXT, Is_Open BOOLEAN, Is_Takeout BOOLEAN)

Users
CREATE TABLE user(UserID VARCHAR(200) PRIMARY KEY, Name TEXT, Review_Count INTEGER, Fans INTEGER, Average_Stars NUMERIC)

Review
CREATE TABLE Review(ReviewID VARCHAR(200) PRIMARY KEY, user_ID VARCHAR(200), business_ID VARCHAR(200), stars NUMERIC, review_date date,FOREIGN KEY (User_ID) REFERENCES Users(ID), FOREIGN KEY (Business_ID) REFERENCES Business(ID))

Address
CREATE TABLE Address(Business_ID VARCHAR(200), City TEXT, State TEXT, Postal_Code TEXT, FOREIGN KEY (Business_ID) REFERENCES Business(ID))

Category
CREATE TABLE Category(ID INTEGER PRIMARY KEY, Type TEXT)

Business_Category
CREATE TABLE Business_category(Business_ID VARCHAR(200), Category_ID SERIAL INTEGER,FOREIG
KEY (Business_ID) REFERENCES Business(ID),FOREIGN KEY (Category_ID) REFERENCES Category(ID))

Rating
CREATE TABLE Rating(Business_ID VARCHAR(200), Star NUMERIC, Review_Count INTEGER,FOREIGN
KEY (Business_ID) REFERENCES Business(ID))

Tip
CREATE TABLE Tip(user_ID VARCHAR(200), Business_ID VARCHAR(200), tip text, Compliment_Count
INTEGER, Tip_Date date,FOREIGN KEY (user_ID) REFERENCES Users(ID),FOREIGN KEY (Business_ID) REFERENCES Business(ID))

4 Dataset reference

https://www.yelp.com/dataset/documentation/main