INTRODUCTION TO BIG DATA
PROJECT: PHASE 3

Pranjal Pandey
Ravikiran Jois Yedur Prabhakar
Venkata Karteek Paladugu
Vishnu Saketh Byreddy

Dataset: Yelp Dataset

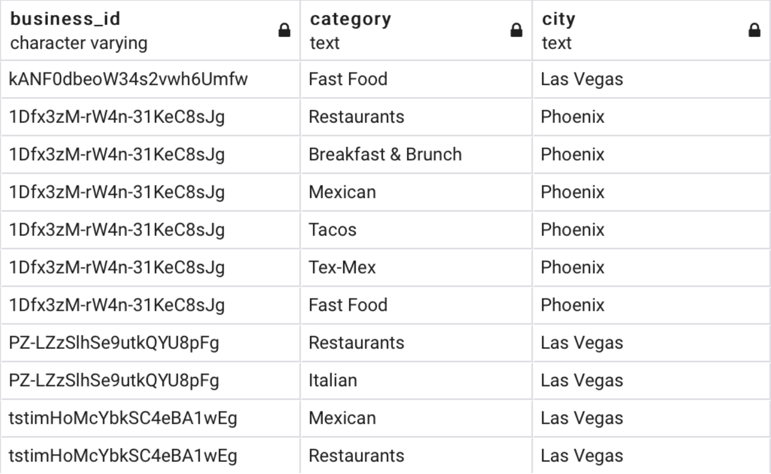
Cleaning and Integration:

The data has been validated and inserted to a newly created table named 'Business_Category_2'. We have filtered the data to be inserted in this table to only include 'Restaurant' category. All the tuples associated with the 'Restaurant' category for a particular city has been inserted to this table. For example, a business has categories such as 'Restaurant', 'Chinese', 'Burgers', 'Sea Food', 'Nightlife' has been inserted under the same business_id so that we can recognize that a single business has this type of cuisine and environment thus, integrating all the information about the restaurant into a single relation.

The category information, which is actually derived from the 'Business' table was in the form of a string with multiple values. This has been converted to array form and then flattened and cleaned to get data that is uniform and also atomic.

The table Business_Category_2 has the following attributes:

- Business ID
- Category
- City the Restaurant belongs to



business_id character varying	category text	city text
kANF0dbeoW34s2vwh6Umfw	Fast Food	Las Vegas
1Dfx3zM-rW4n-31KeC8sJg	Restaurants	Phoenix
1Dfx3zM-rW4n-31KeC8sJg	Breakfast & Brunch	Phoenix
1Dfx3zM-rW4n-31KeC8sJg	Mexican	Phoenix
1Dfx3zM-rW4n-31KeC8sJg	Tacos	Phoenix
1Dfx3zM-rW4n-31KeC8sJg	Tex-Mex	Phoenix
1Dfx3zM-rW4n-31KeC8sJg	Fast Food	Phoenix
PZ-LZzSlhSe9utkQYU8pFg	Restaurants	Las Vegas
PZ-LZzSlhSe9utkQYU8pFg	Italian	Las Vegas
tstimHoMcYbkSC4eBA1wEg	Mexican	Las Vegas
tstimHoMcYbkSC4eBA1wEg	Restaurants	Las Vegas

Itemset Mining:

The main inspiration for the frequent itemset mining was to find the itemsets containing environment and types of cuisines offered in restaurants across some of the popular cities in the United States. The cities that were considered were: Washington, Dallas, Las Vegas, Denver and Phoenix. This information can be helpful to potential investors so that they will know what kind of restaurants to invest in and where to invest in, thus, increasing the potential for growth and profit.

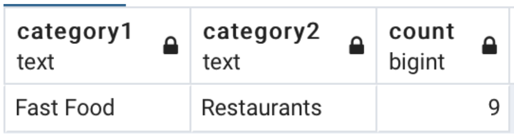
We used Apriori algorithm to do the frequent itemset mining. We created lattices for each city based on a minimum support for each city.

The minimum support assigned for each city is as follows:

Dallas: 5, Las Vegas: 40, Phoenix: 40 and Denver: 5

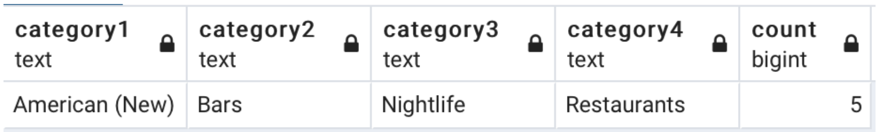
We got the following results as the top level of the lattice for each city:

- *Dallas*:



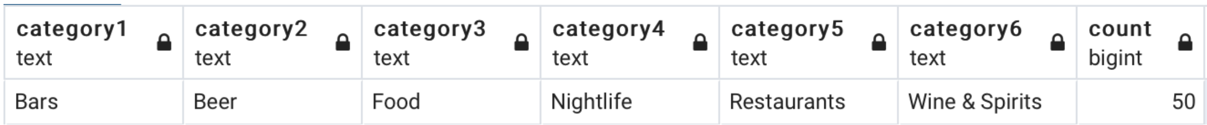
category1 text	category2 text	count bigint
Fast Food	Restaurants	9

- *Denver*:



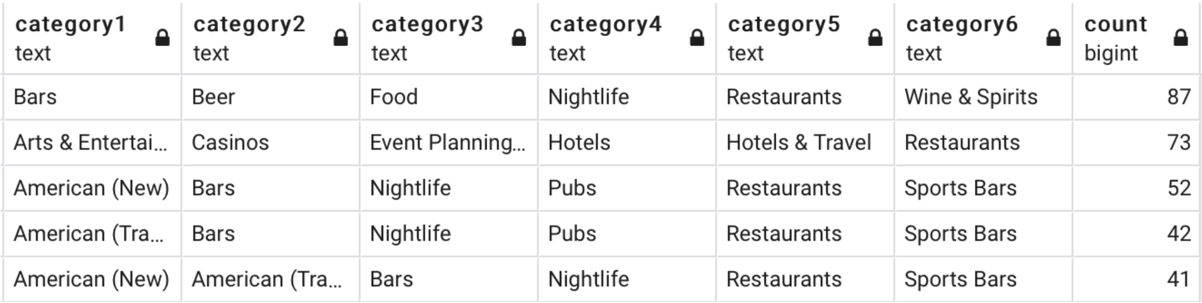
category1 text	category2 text	category3 text	category4 text	count bigint
American (New)	Bars	Nightlife	Restaurants	5

- *Phoenix*:



category1 text	category2 text	category3 text	category4 text	category5 text	category6 text	count bigint
Bars	Beer	Food	Nightlife	Restaurants	Wine & Spirits	50

- *Las Vegas*:



category1 text	category2 text	category3 text	category4 text	category5 text	category6 text	count bigint
Bars	Beer	Food	Nightlife	Restaurants	Wine & Spirits	87
Arts & Entertai…	Casinos	Event Planning…	Hotels	Hotels & Travel	Restaurants	73
American (New)	Bars	Nightlife	Pubs	Restaurants	Sports Bars	52
American (Tra…	Bars	Nightlife	Pubs	Restaurants	Sports Bars	42
American (New)	American (Tra…	Bars	Nightlife	Restaurants	Sports Bars	41

From the above results, we can easily deduce that the restaurants in Las Vegas mostly give the following services: Pubs, Bars, Beer, Food, Nightlife, Wine & Spirits and Sports Bars.

In all other cities that we have considered (not excluding Vegas), we can see that most of the restaurants provide Bars and Nightlife services. From this, we can deduce that it is a safe option to invest in these categories.

Association Rules:

We have written a code to find interesting association rules for each city that is considered. We have set the minimum confidence as 0.4 and lift as 1.

Some of the association rules that we have found are as follows:

- *Dallas*:
  No association rule has been found to be interesting as the amount of data found in the lattice structure is very less

- *Denver*:
    - {‘Arts & Entertainment', 'Restaurants', 'Hotels', 'Hotels & Travel', 'Casinos'} --> {'Event Planning & Services'}
        - The lift value is: 20.395
        - The confidence value is: 1.0
        - This suggests that the restaurants which have Hotels, Hotels & Travel services and Casinos as their categories also provide Event Planning & Services

    - {'Pubs', 'Bars', 'Sports Bars', 'Nightlife', 'Restaurants'} --> {'American (New)'}
        - The lift value is: 6.887
        - The confidence value is: 0.702
        - This suggests that the restaurants which have Pubs, Bars, Sports Bars and Nightlife categories also provide American (New) cuisine

- *Phoenix:*
    - {'Sports Bars', 'Restaurants', ‘Bars', 'Nightlife'} --> {'American (Traditional)'}
        - Confidence: 0.650
        - Lift: 4.939
        - This suggests that the restaurants which provide Bars, Nightlife, Food and Sports Bars also provided American (Traditional) cuisine

    - {'Beer', 'Bars', 'Restaurants', 'Nightlife', 'Food'} --> {'Wine & Spirits'}
        - Confidence: 0.872
        - Lift: 5.989
        - This suggests that the restaurants which provide Beer, Bars, Nightlife and Food also provide Wine & Spirits

- *Las Vegas:*
    - {'Restaurants', 'Burgers', 'Breakfast & Brunch'} --> {'American (Traditional)'}
        - Confidence: 0.652
        - Lift: 4.701
        - This suggests that the restaurants which provide Burgers, Breakfast & Brunch provide American (Traditional) food
    - {'Event Planning & Services', 'Venues & Event Spaces'} --> {'Restaurants'}
        - Confidence: 1.230
        - Lift: 1.230
        - This suggests that the businesses which provide Event Planning & Services and Venues & Event Spaces are Restaurants too.

We believe that the Relational database model is more suitable for itemset mining and data integration because, it has a specific schema that we can follow and the queries to be written for the Apriori algorithm is much more developer friendly than writing the query for the Document based model i.e., MongoDB.

However, cleaning the data in MongoDB is easier as we get access to each and every document separately. MongoDB also has an easier way to access and alter information stored in the arrays.

The read and write speeds in PostgreSQL is faster than that of MongoDB [1]. Joining relations is faster in PostgreSQL. Since data integration involves many joins across tables, we have chosen PostgreSQL.

Reference:

[1] Big Data Performance and Comparison with Different DB Systems
http://ijcsit.com/docs/Volume%208/vol8issue1/ijcsit2017080114.pdf