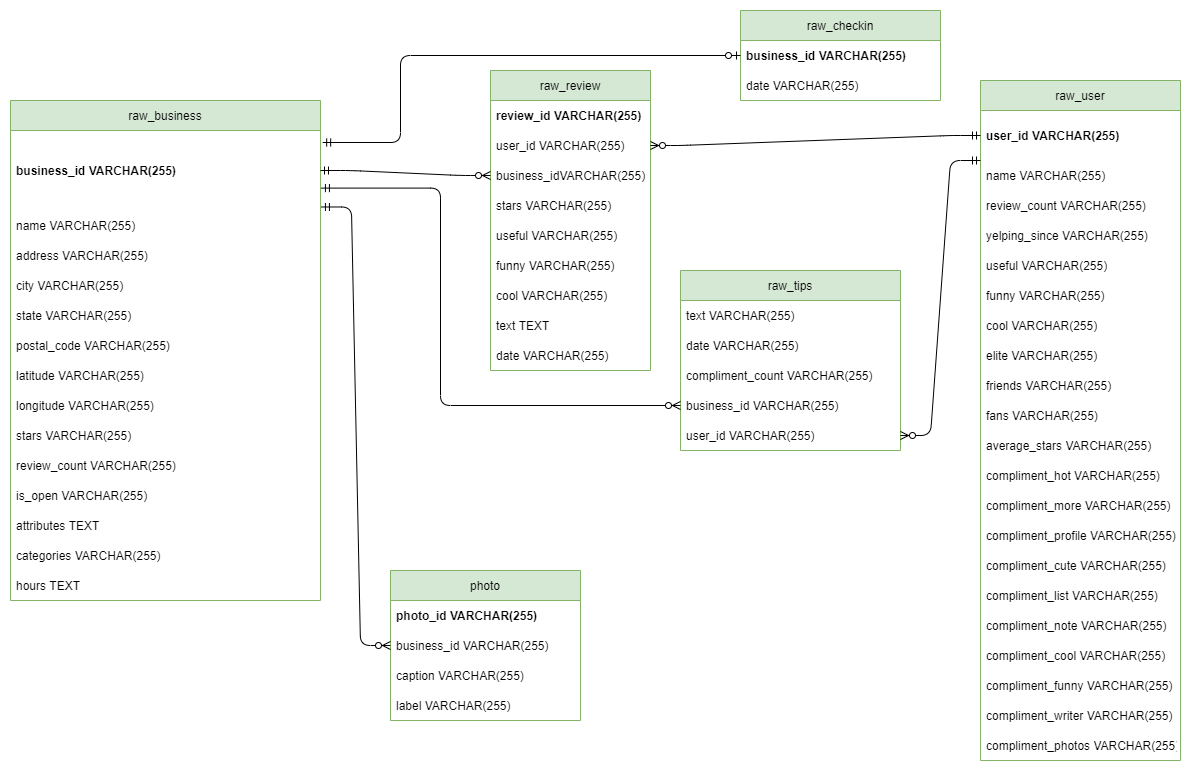
**YELP DOCUMENTATION**

Firstly I created the default schema and relationship by the documentation provided by the YELP.

Here, I have not used any kind of constraints.



*Fig:ER diagram of Raw tables*

Logical Modelling:

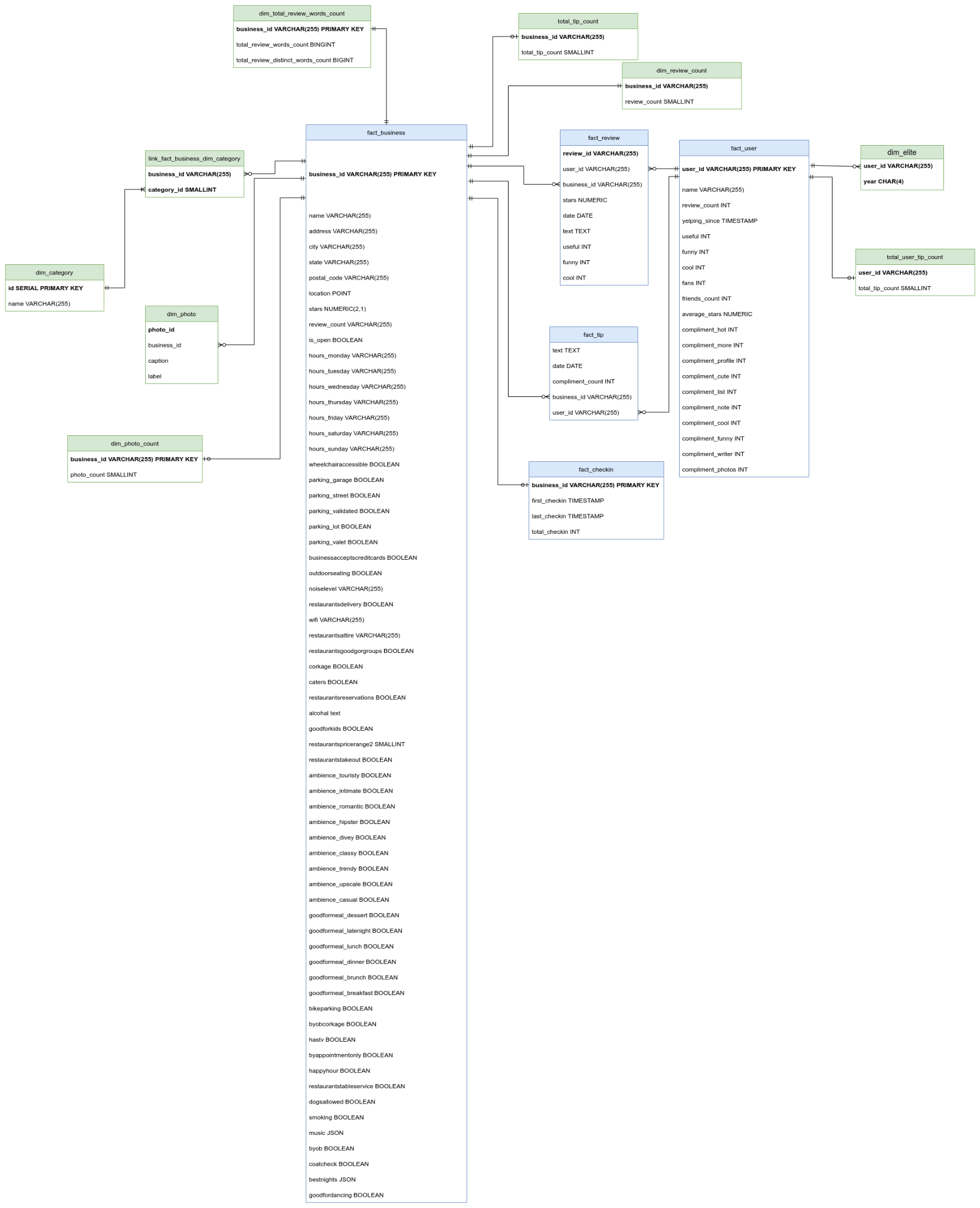
After the data was pushed into the raw table, I had a lot of insight about the data and I started further exploring its entities and attributes and the data types for it.

Below is the list of entities , description and domain of the model.

|  |  |  |
| --- | --- | --- |
| Entity | Description | Domain |
| dim\_category | The categories of the business |  |
| Attributes:  id  name | Identifier for the entity,SK,FK  The name of the category | Auto generated,Serial  Text |
| link\_fact\_business\_dim\_categoty | The table which link entity fact\_business and dim\_cateogty which has the many to many relationship with each other |  |
| Attributes:  business\_id  category\_id | Identifier of the entity fact\_business, FK  Identifier of the entity category | Id which references the table fact\_business  Valid Id which references the table dim\_cateogty |
| dim\_photo | The photos id of the business |  |
| Attributes:  photo\_id  business\_id  caption  label | The identifier which represents the photo  FK which references the fact\_business  The caption of the photo  The label of the photo | Valid business\_id from table fact\_business  Text  Text |
| dim\_photo\_count | The counts of photos associated with a business. |  |
| Attributes:  business\_id  photo\_count | PK and FK  The count of photos of a specific business | Valid Id from table fact\_business  Natural Number |
| fact\_business | The details about the business. |  |
| Attributes:  business\_id  name  address  city  state  postal\_code  location  stars  review\_count  is\_open  hours\_monday  hours\_tuesday  hours\_wednesday  hours\_thursday  hours\_friday  hours\_saturday  hours\_sunday  wheelchairaccessible  parking\_garage  parking\_street  parking\_validation  parking\_lot  parking\_valet  businessacceptscreditcards  outdoorseating  noiselevel  restaurantsdelivery  wifi  restaurantattire  restaurantsgoodforgroups  corkage  caters  restaurantsresercations  alcohal  goodforkids  restauranrspricerange2  restaurantstakeout  ambience\_touristy  ambience\_intimate  ambience\_romantic  ambience\_hipster  ambience\_divey  ambience\_classy  ambience\_trendy  ambience\_upscale  ambience\_casual  goodformeal\_dessert  goodformeal\_latenight  goodformeal\_lunch  goodformeal\_dinner  goodformeal\_bunch  goodformeal\_breakfast  bikeaparking  byobcorkage  hastv  byappointmentonly  happyhour  restaurantstableservice  dogsallowed  smoking  music  byob  coatcheck  bestnights  goodfordancing | Identifier of the entity , PK  The name of the business  The address of the business  The city in which the business is located  The state in which the business is located  The postal code of that area  Latitude,Longitude of the business,GPS  The star rating,rounded to half-stars  The total count of the reviews  Boolean, True or False for closed or open  Start\_time and end\_time separated by dash  Start\_time and end\_time separated by dash  Start\_time and end\_time separated by dash  Start\_time and end\_time separated by dash  Start\_time and end\_time separated by dash  Start\_time and end\_time separated by dash  Start\_time and end\_time separated by dash  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Text Value telling the noise  Boolean value showing true or false  Boolean value showing true or false  The text value showing the attire  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Text value telling the info about alcohol  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  JSON showing the music keys and values,Taking this as JSON as only 7,198 values are not null  Boolean value showing true or false  Boolean value showing true or false  Boolean value showing true or false  JSON showing the music keys and values,Taking this as JSON as only 5,526 values are not null | 22 Character  Text  Text  Text  State Code  Text  GPS,POINT  Numeric  Int  Boolean  Text  Text  Text  Text  Text  Text  Text  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Text  Boolean  Boolean  Text  Boolean  Boolean  Boolean  Text  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  Boolean  JSON  Boolean  BooleanBoolean  Boolean  JSON |
| dim\_total\_review\_words\_count | The table showing the total review words count and the total review distinct words count |  |
| Attributes:  business\_id  total\_review\_word\_count  total\_review\_distinct\_words\_count | Identifier which represents the business, PK,FK  The total word counts of all reviews  The total distinct words counts of all reviews of a business | Valid ID from table fact\_business.  INT  INT |
| total\_tip\_count | The total tip count of the business |  |
| Attributes:  business\_id  total\_tip\_count | Identifier which represents the business,PK,FK  The total tip count of a business | Valid ID from the table fact\_business  INT |
| fact\_review | The details about the reviews made by the user to the business |  |
| Attributes:  review\_id  user\_id  business\_id  stars  date  text  useful  funny  cool | Identifier of the entity,PK  Identifier of the entity user,FK  Identifier of the entity business,FK  The stars given by the user to the business  The date at which the review was given  The review text  The number of useful reaction by users  The number of funny reaction by users  The number of cool reaction by users | Text  Text  Text  Numeric  Date  Text  INT  INT  INT |
| fact\_tip | The tip given by the user to the business. |  |
| Attributes:  text  date  compliment\_count  business\_id  user\_id | The text about the tip given to the business by the user  The date at which the tip was given  The count of the compliment  The business to which the tip is given  The user who gives the tips . | Text  Date  INT  Valid ID from table fact\_business  Valid ID from table fact\_user |
| fact\_checkin | The checkin done on the business. |  |
| Attributes:  business\_id  first\_checkin  last\_checkin  total\_checkin | The identifier which references the business,FK  The TIMESTAMP at which the first checkin  was done  The TIMESTAMP at which the last checkin  was done  The total number of checkin done | Valid ID from the table fact\_business  TIMESTAMP  TIMESTAMP  INT |
| fact\_user | The info about the user |  |
| Attributes:  user\_id  name  review\_count  useful  funny  cool  fans  friends\_count  average\_stars  compliment\_hot  compliment\_more  compliment\_cute  compliment\_list  compliment\_note  compliment\_cool  compliment\_funny  Compliment\_writer  compliment\_photos | Identifier of the entity fact\_user,PK  The name of the user  The total review count of the user  The total number of useful reaction received to their reviews  The total number of funny reaction received to their reviews  The total number of cool reaction received to their reviews  The total number of fans of the user  The total number of friends of the user  The number of compliments\_hot  The number of compliments\_hot  The number of compliment\_more  The number of compliment\_cute  The number of compliment\_list  The number of compliment\_note  The number of compliment\_cool  The number of compliment\_funny  The number of Compliment\_writer  The number of compliment\_photos | PK  Text  INT  INT  INT  INT  INT  NUMERIC  INT  INT  INT  INT  INT  INT  INT  INT  INT  INT |
| dim\_elite | The year in which the user was elite. |  |
| Attributes:  user\_id  year | Identifier which references the user.PK  The valid year at which the user was elite,PK | Valid ID referencing the fact\_user table  CHAR |
| total\_user\_tip\_count |  |  |
| Attributes:  user\_id  total\_tip\_count | Identifier which references the user.PK  The total number of tips count by user | Valid ID from fact\_user table.  INT |

Proposed ER model

Below is the proposed ER diagram of the warehouse.



Validation:

Fan-Trap

It looks like there is a fan trap everywhere throughout the ER model, but the above fan trap is not going to affect our model as per my design.

Chasm-Trap

Since, the pathway exists between all of the entities, so there's no occurrence of the chasm-trap.

Physical Implementation:

1. Making raw tables
2. Pushing the data from the pipeline into the raw tables.
3. Data cleaning of the table.

* Making the ‘None’ value as NULL.
* Changing the data types of the attributes as per the model proposed.
* Changing the values like ‘no’ , ‘uno’, uyes\_corkage’,’uyes\_free’,’yes\_free’ to 0 and 1 and casting it to Boolean.
* While splitting the words by comma on the category field of the business entity, spaced at the beginning was trimmed off to make consistency among the same category name.
* Making the elite years ‘20,20’ as the 2020.

1. Creating the schemas as proposed above in the ER model.
2. Pushing the data into the fact and dimension table by further cleaning the data if necessary.
3. Validation of different aspects of the data, such as

* Checking the total photo\_id and the total distinct photo\_id is 0.
* Checking all the unique photos are associated with the unique business\_id
* Checking if all the friends are not the user
* Checking if the yelping\_since is not in the future
* Checking if the average\_stars is not in between 0 and 5
* The review count of a business is not equal to the provided reviews\_count at fact\_businesss.

1. Exporting the data into flat files for the visualization of data on power BI as I have DBMS on the linux system on my computer, and PowerBI is only supported on the windows.
2. Finally uploading the data on PowerBI for visualization.

CODE DESCRIPTION LINK:

<https://github.com/callingsandesh/yelp/blob/final_project/docs/code_explanation.md>

Visualization:

LINK:

<https://github.com/callingsandesh/yelp/blob/final_project/docs/Dashboard%20and%20Report.pdf>