**Normalization Process**

At this point in the database creation process, the tables that have been created thus far were to be normalized. This required ensuring each data table follows the three Normal forms as outlined below:

1. First Normal Form – To satisfy first normal form each attribute cannot contain a list of multiple items. Each column must also contain the same type of information and not vary in datatypes. Additionally, the columns within a table were to have identifiable column names and no repeats. The final condition of first normal form is to ensure that data is stored in a manor where order does not affect it.

1. Second Normal Form – To satisfy second normal form each table must first be in first normal form. Furthermore, each table cannot have any partial dependencies essentially meaning that each attribute must be dependent on the entirety of the primary key whether or not it is composite or not. Finally, there can be no calculated fields like age in the data.
2. Third Normal Form – To satisfy third normal form the database must meet First and Second Normal form conditions. Then, the database must have no transitive dependencies which essentially means none of the columns can depend on another column which is not the primary key

In this step of the database creation, the following changes were made to ensure that all 3 Normal Forms were satisfied.

FIRST NORMAL FORM

1. The Job postings table had one column titles details which was an accumulated list of all requirements and bullets during scraping. This was the method of scraping because the html of the site was identical for each bullet and could not be extracted any other way. The details list was removed from the main table and made into its own job\_requirements table. The lists were unpacked, and each bullet was made its own row with a one to many reference to the job\_ID in the job\_posting table.

**Sample Code for this change**

job\_requirements\_df = pd.DataFrame(job\_posting\_teamwork\_df[['job\_ID','details']])

job\_requirements\_df\_final = job\_requirements\_df.assign(temp = job\_requirements\_df.details.str.split(",")).explode('details').drop('temp',axis=1)

job\_requirements\_df\_final['details'] = job\_requirements\_df\_final['details'].str.replace("'","").str.replace('"','')

Graphical user interface, text, application, email

Description automatically generated

1. The job location within the job\_posting table, the headquarters field in the company\_team table and tweet\_location were all formatted as ‘City, State Abbreviation’. To normalize this further the location and headquarter fields were split into a city field and state field.

**Sample Code for this change**

job\_posting\_teamwork['job\_city'] = job\_posting\_teamwork['Location'].str.partition(",")[0]

job\_posting\_teamwork['job\_state'] = job\_posting\_teamwork['Location'].str.partition(",")[2]

Table

Description automatically generated with low confidence

Table

Description automatically generated with low confidence

1. Teamwork Online does not provide salary information nor application deadlines so the fields needed to be passed as NULLS. This was done by omitting the fields during the insert statement. However, some fields did have values for some records and not for others including numeric fields like capacity and founded\_year in the company\_team table. These were passed with a dummy value like NA or 0 for numerics. In order to ensure the right data was passed and the type was the same for each individual column Table Updates were completed to set them to NULLS.

**Some Sample Code for this change**

UPDATE company\_team

SET league = NULL

WHERE league = 'NA';

UPDATE company\_team

SET league\_short = NULL

WHERE league\_short = 'NA';

UPDATE company\_team

SET stadium = NULL

WHERE stadium = 'NA';

UPDATE company\_team

SET stadium\_capacity = NULL

WHERE stadium\_capacity = 0;

UPDATE company\_team

SET founded\_year = NULL

WHERE founded\_year = 0;

SECOND NORMAL FORM

1. To ensure there were no calculated fields, only datestamps were used including a column called scrapedatetime to provide a reference time for the Teamwork online data which had no application info. This field and other dates were kept in raw form and not used for any calculations.
2. Final Second Normal Form check was completed to ensure that there were no partial dependencies. After analyzing each table it was determined they had no partial dependencies.

THIRD NORMAL FORM

1. A check was run to ensure there were no transitive dependencies or dependencies on non-primary key attributes. Initially it appeared that maybe job\_title had a transitive dependency in job\_posting however it was observed that the attribtues in fact do not depend on job\_title because they may be and in fact are different for jobs with the same title and different job\_ID. Thus this proves the tables are in Third Normal Form.

**Some Final Tables to illustrate Normalization:**

Graphical user interface

Description automatically generated with medium confidence

Table

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated