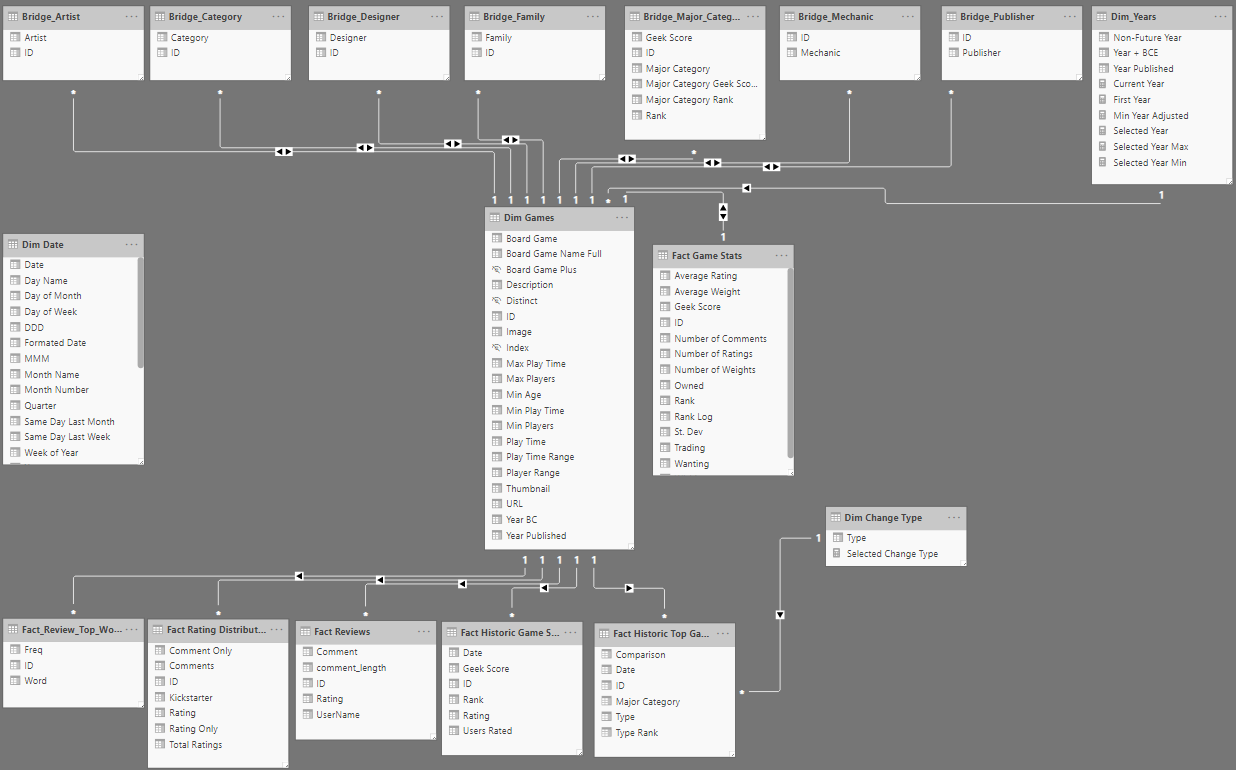
**Building the Board Game Geek Dashboard**

**Inception:**

This project has been something that I have been wanting to do for a long time, but I did not have the experience to put it all together. When Decisive Data gave me the opportunity to pick any topic to build a dashboard, I naturally leapt at the chance to build something about boardgames. This naturally led me to the Board Game Geek API. At the time that I had started I had no experience with pulling data directly from an API and so I settled for a static set of data I was able to pull from Kaggle. As I worked with the data and began to learn more through my role at Decisive Data, I learned about Power BI’s ability to pull data directly into using an API.

**The Data Model:**



**Part 1: Dim Games, Fact Game Stats, Bridge Tables, and Dim Years**

To create my data model, I first connected Power BI directly to Board Game Geek’s XML API. Using the built in XML parser in Power BI, I separated out the many to one tables to create a connection between each game and its Artists, Designers, Publishers etc. and labeled these tables with the Bridge Prefix. Traditionally, the bridge tables would connect to normalized dimensions (ie. Dim Artists), however, Power BI is optimized to deduplicate data and creating these dimensions tables is unnecessary.

These tables were created by pulling in the base data and then separating it out into each of the individual tables via referencing. Dim Years was created using the distinct list of years from the year published in Dim Games. Dim Years and Fact Game Stats could be combined into the Dim Games table. I chose to keep them separate as they represent different information. Keeping them separate allows a user to clearly see the purpose of each table and leads to a model that is more easily understood.

**Part 2: Fact Reviews, Fact Rating Distribution, and Fact Review Top Words**

Unlike the first set of tables, this data could not be pulled directly from the BGG API and required a static file of reviews. This data does not change over time and all the processing to create it was done in R before being imported into Power BI. The review data comes from a set of ~13 million reviews that were pulled on May 2nd, 2019. This one set of data was split into three tables:

The reviews consisted of a game ID, username, rating, and comment. To prep this data, the ratings were rounded to the nearest whole number. The ratings where then split into 3 groups: ratings without comments, ratings with comments that reference Kickstarter, and ratings with comments that did not reference Kickstarter. The ratings were then grouped by the rating and game ID to create Fact Rating Distribution. This is then used to create the

The Review Top Words table was created using the comments for each game and removing stop words, the board game name, and common words in all reviews (such as game and players). For full list of removed words see the appendix. The top 30 words for each game were then calculated and uploaded to Power BI.

The Fact Reviews table

**Appendix:**

**Words Removed:**

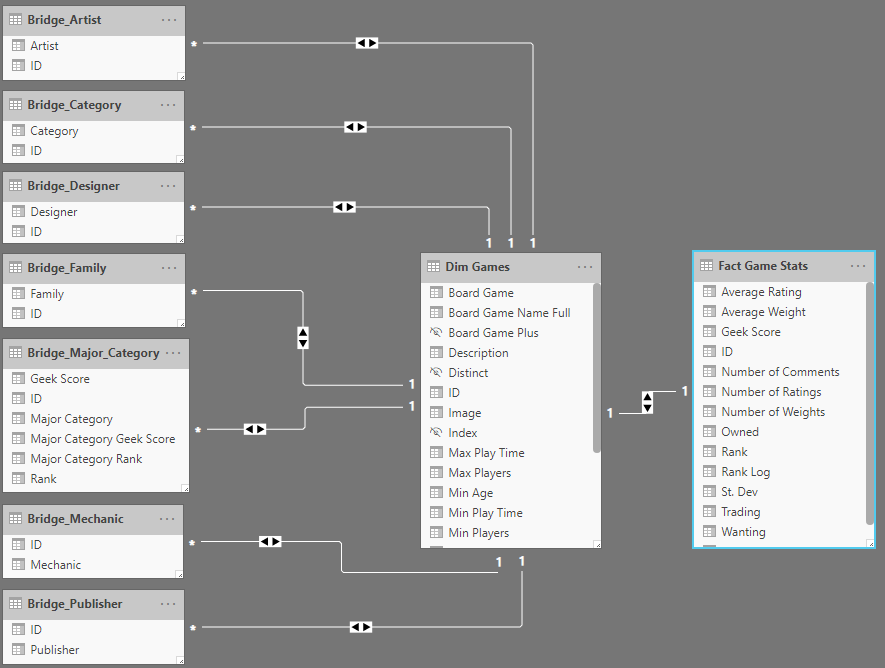
The following words were removed specifically from the reviews: game, games, play, played, the, board, one, can, still, get, this, just, will, its, got, lot, now, player, playing, players, many, always, ive, times, never, bit, plays, also, though, every, gamers, but, really, dont, even, like, much, well, way, make, opinion, enough, version, quite, another, end, actually, see, need, box, set, take, around, doesnt, based, seems, didnt, right, cant, find, want, makes, que, ever, almost, know, try, add, bought, bought, use, lots, copy, idea, may, buy, probably, might, although, less, going, nothing, away, made, feels, gets, number, theres, however, something, rather, give, anything, juego, adds, definitely, things, sure, table, able, real, rule, year, years, amount, since, last, base, keep, come, found, short, either, basically, isnt, place, ago, without, bits, addition, trying, looking, added, least, used, especially, part, aspect, anyone, para, already, look, push, words, youre, los, early, along, taking, day, similar, takes, comes, change, arent, jeu, instead, done, asi.

Note: Punctuation was removed from all words.

**The Data Prep:**

Once I connected Power BI into the API, I was able to begin pulling the data. The API pull from Board Game Geek uses XML which Power BI parses natively. There were several optional queries that can be appended to the end of the API call. I started first as pulling the data for a single game and building out a database. Using Power BI’s XML parser made navigating the data easy and by referencing the data, I was able to build out a snowflake schema quickly.

The snowflake schema is optimal for Power BI data models as Power BI is optimized to deduplicate data and does not require a fully normalized data model. There are some instances in which you will want to normalize your data, however, in most cases this is unnecessary and only adds more processing time. The initial data model I created can be seen below.



With the data model created, it was time to turn to pulling in more records. The first part was to find a list of game IDs to pass into the API. The initial list I had was a good starting point and by splitting the full list into smaller groups I was able pass multiple IDs at a time. This led to a few problems as the BGG API restricts the frequency of queries from a single IP address. To avoid over querying the data, I added in a wait function to add a few seconds between each query. While this increased the time to pull the data is ensured I never hit the frequency limit and had my data pull fail.

The second problem was that the list of game IDs I had was static and wouldn’t pull new games. After some searching, I found a fellow board game enthusiast who had a GitHub repository where he stored historic game statistics, including a daily updated set of game IDs. Using Power BI I dynamically referenced his repository to pull in the latest days list and pass the game IDs through the BGG API to pull update full game information daily. With this new source I had the data model that I originally sought to build.

I quickly realized that I had stumbled upon a trove of information that I could use to further improve my dashboard. I was no longer limited to the most recent data pull. There was also the matter of the 7 million reviews that I had not yet touched. Thus, began the second iteration of the dashboard.

To fully utilize the new information, I turned to R to help parse the data and prep it for Power BI. The first