Phase 2

Jason Li, Guanyu Song

November 20, 2021

Design Decisions

We decided to create a new table called Asteroid for moon like objects which no parent planet. We chose to create a new asteroid table instead of merging it into the moon table is so that we have as few null values in the tables as possible. Since the different categories of objects have different recorded and calculated values, some will have certain null attributes while others will not. Having more tables, rather than fewer, will result in us having tables as full as possible.

We changed the attribute name of the id's in the tables to be all uppercase to better match the style in assignment 2. We decided capitalizing the id's like in assignment 2 was a good style choice because we can easily see which attributes are id's and which are not.

We decided to remove the default value category of a planet. Instead, we changed it to add a NOT NULL constraint. We decided it made little sense to categorize a planet as full sized or dwarf if we did not know how to. We added the NOT NULL constraint for the planets in the database, we can categorize all of them because of their vicinity to earth and how much research we have on them.

We changed the primary key of the Discovered table to be the celestial body as well as the astronomer. This is because it can be the case multiple people collaborated to help discover the same celestial body.

For the string like attributes in the database, we debated on whether to have them be a varchar or text. We decided to have the attributes be TEXT type because it did not make sense to put a cap on the length of the string since names of planets and people can get longer than expected. Furthermore, there is not much improvement to memory or speed if the change to varchar is made.

We changed the isMoon table to Orbit. We did this because we wanted to add orbit information such as perihelion and aphelion which does not solely belong to the satellite or parent body. It also did not make sense to have separate isMoon and orbit tables. Finally, this choice made it so that this table can be easily extended to include orbits that are not just moon and planet. Orbit can also have rows in the table representing the orbit of a planet around its parent star.

We deleted the moonNumber attribute in Planet because the number can be found with a simple count query.

Cleaning Process We tried to minimize the clean up needed by changing the schema in ways described in the design decisions section.

We had to make changes in the purposed schema to better match the data we are using. We changed attributes type and escape in the Planet table to new keywords. This is since type and escape are reserved keywords in postgresql. We changed them to new attributes, 'category' and 'escape_speed'.

For the dates in the csv file from the data source, the time stamps were in many different formats. For example dd/mm/yyyy or yyyy/mm/dd. Some time stamps also only had did not have a specific day or month. This was especially common for older time stamps where the specific date of events is less well documented. To insert the time stamps with different formats, we inserted them separately for different time stamp formats. When the day or month was missing, we just added the year or year and month without the missing piece(s) of information.