CSC343 Phase 3

Data Cleaning Steps

1. Sort the rows by alphabetically and numerically sorting the entity_id. This ensures that when we delete the rows, we keep a consistent range of entity_ids within all of our tables. For example, if in table1, we have values from “c1” to “c1000”, in table2, we want values from “c1” to “c1000”, instead of any values from ranges “c1001” and onward. This is important because our entity_id is a referential integrity constraint for many of our tables.
2. Delete any rows with NULL values or values we don’t need. For example, if the valuation amount is $0 in our IPOS table, then we deleted this row since it would affect any aggregation functions we would do in the future.
3. Delete rows that would overload the server. We kept at least 1000 rows in each of our tables (except for the tables that had less than 1000 rows of data), and deleted the rest of the rows. For the Startups table, we kept 2000 rows because it contains entity_id, which is a very important key constraint for many other tables.

Schema Changes

While we were cleaning the data, we noticed some inconsistencies in the data that led us to alter our schema. We:

- Changed entity_id from an integer to a character.
- In Investments, changed primary key to be entity_id, funding_round_id, and investor_object_id, since multiple startups can go through multiple funding rounds
- Changed person_id to an integer from a character

Decisions

Various decisions were made to ensure our data is clear. These include:

- Deleting all the valuation_amounts in the IPO table or raised_amount in the Investments table that are $0. The rows that contain a $0 dollar amount is similar to NULL, and our questions did not need any of this information.
- Deleting the extra rows that would overload the server. For example, in table Investments, we only kept 1000 rows, and deleted the rest. If in Phase 4, we do not have adequate information to answer the questions, we will import more data.