Milestone 1

A brief description of your Application:

For our project, we are going to scrape SEC quarterly reports of companies, process and store the data. We are then going to make a web application to query the database and make company and industry wide inferences on the data. The goal is to offer users a way to get the information they need on companies for investment decisions in a simple, clear, and standardized format.

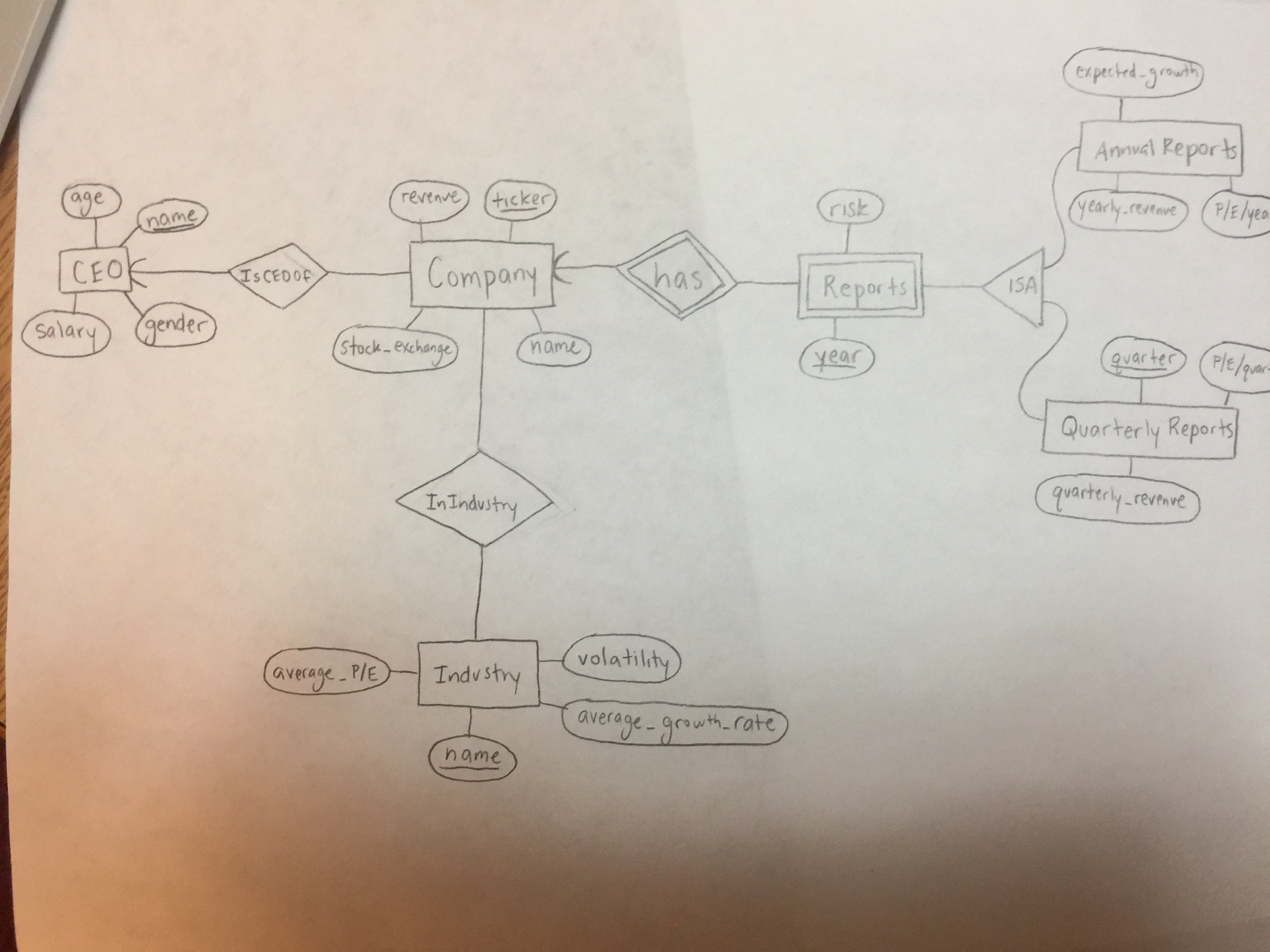
Plan for populating the database:

Public companies must publish quarterly reports with the SEC. These reports are publically available on the SEC’s website. Since the files are stored in XML format we can parse the data and upload it to our database. The tags and format of each company’s xml structure will vary slightly between companies. Depending on the size of all the disparities we will implement a methodology for hashing the document tags to our database columns either by coding in the column names, considering the surrounding context/tags, using machine learning. Until we get this setup we will just manually input sample data.

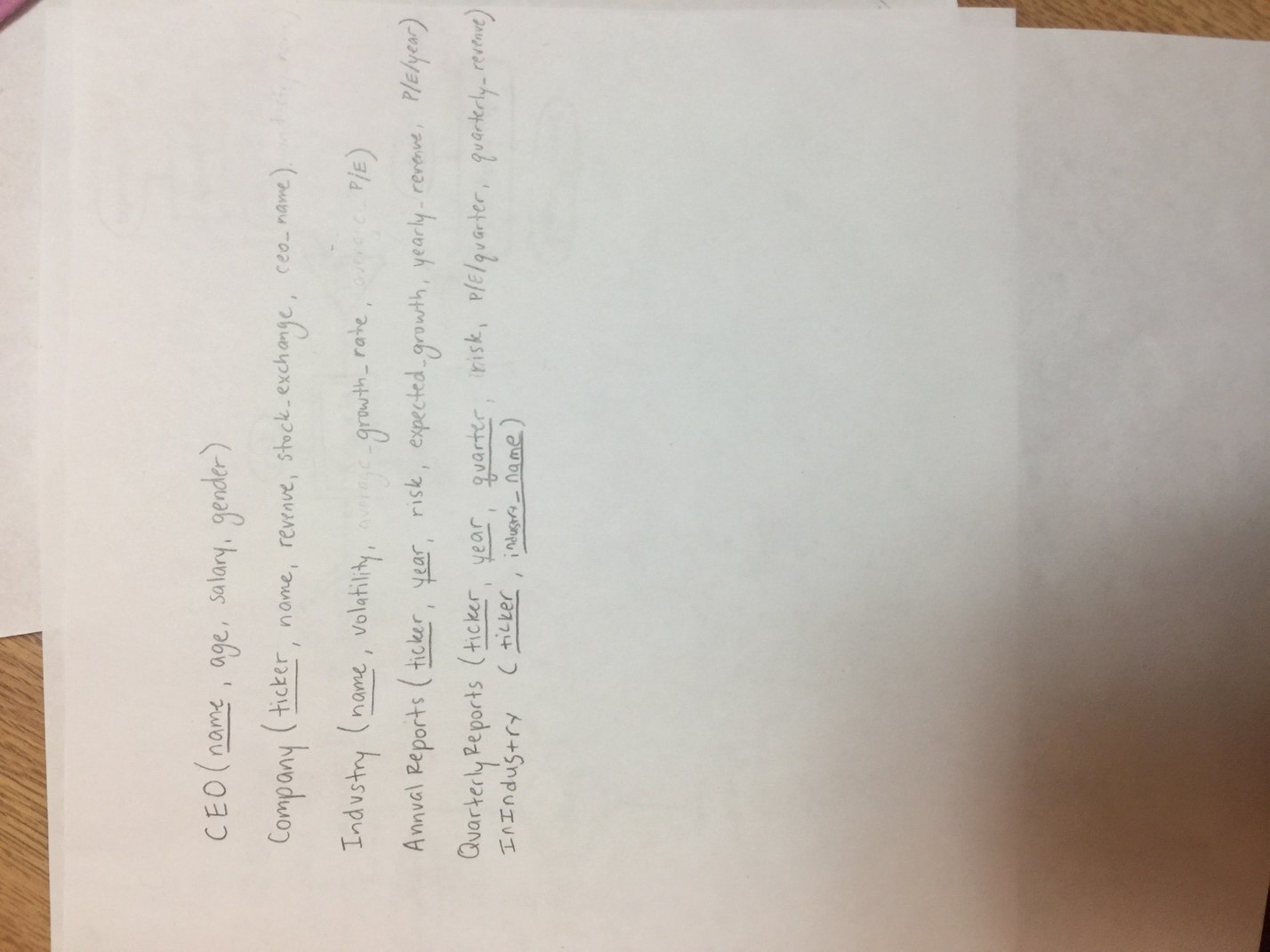
Assumptions:

Since we are making the ticker for a company the key for company entity, we assume that every company that we scrape has a unique ticker symbol and that is also not null. Furthermore, we can assume any key attribute will have a non null unique value when we scrape the data as this is a rule for primary keys. As of right now, our databases assume everything is not null but this may change in the future. Another assumption we made is that every company has exactly one CEO. With this assumption, we can get rid of the relationship table IsCEOOf. We assumed that a company can be in multiple industries as companies cross different fields. This requires us to have a relationship table in our schema. We also assume that based off the data we scrape, we can gather certain information about the respective entities. We assume that we can obtain the gender of a CEO and also what specific industry a respective company is in.

E/R Diagram:



Schema/List of Tables:



Web Interface:

The user will be have fields to search by company ticker to get a specific company (with a text box), or to narrow down to companies which match certain parameters. Parameters could include ranges (with text boxes) for year, earnings, expected growth, and other worthwhile metrics. Once the user has specified his desired parameters, he would receive a list of reports that match the parameters and be prompted to select (with checkboxes) which companies to generate the reports for and (from dropdown boxes) which metrics to be included. Then a zipped folder with the reports (in a standard format like .xls or .csv) would be returned as a download link.