Final Report: Amazon Phone Review ETL Project

Matt Andrews, Gregory Forbes, Pete Farquhar

11/13/2019

Extracting:

We chose to work with two different datasets (3 files total) from Kaggle that were large CSV files describing mobile phone reviews on Amazon. Two of the files were from one source, with one file containing descriptive fields of phones and the other file a long list of 80,000+ reviews. The other file was a similar long list of 4,000+ reviews.

Transforming:

Our first task was to merge the two CSVs that came from the same source, using a jupyter notebook and Pandas, which was straightforward because they had a shared index. This matched the descriptive data on the phones with the long list of reviews.

We were less successful in our second task. Our hope was to match some of the reviews from the third dataset to some of the same phones from the first two datasets. We tried to merge by matching product names, the url for the product image, and the url from the product review. None of these were successful. The product names were too long and unique as strings to match each other. For the urls, we discovered that the third dataset was from India, so the urls were all “amazon.in” rather than “amazon.com”. As a result, we just went ahead with the merged file of the first two datasets. We renamed some columns to make the data cleaner overall, and exported as a .CSV.

Loading:

Loading was also time consuming, due largely to the presence of long strings values—the body text of the reviews. We created a Schema in Postgres SQL and attempted to import the file. Postgres had difficulty with the long string values, because of both commas and quotation marks in the fields. As a result we had to drop the longer string fields.

Conclusion:

The whole project was more work than we anticipated due to the “messiness” of the data. This presented challenges at each stage, so that our attempts to merge and upload the data took many attempts as we worked with small glitches. Given more time we would have manipulated the string data to make things run more smoothly. In addition, we could have just appended the Indian reviews using a few common fields.