Final Report for ETL:

Deepak V. & Robert S.

Deepak and I used Google’s Cloud Platform to complete this project. First things first, we found our dataset on Kaggle. It included 2 large csv files that contained Marvel and DC superhero information. This data was provided by FiveThirtyEight and was scraped from Marvel’s and DC’s wiki pages. The data set is maintained using GitHub’s and Kaggle’s APIs.

Google’s Cloud Platform allowed us to use one their servers to Extract and Load our data to their SQL server. In order to do so we had to create an instance on their server, then connect to it (sounds a lot easier than it was). After connecting to it, we loaded in our 2 super csv documents. After we loaded in our csv’s to the GCP, all we had to do was transform and clean the data a little. To do that we set up specific columns that we wanted to load the data into. When the data was cleaned and loaded into the SQL server on Deepak’s computer, we used SQL workbench on Rob’s computer to connect to that instance by telling GCP to allow access to Rob’s IP address.

This was different way to complete this project, but it allowed us to get experience using GCP. Many businesses use GCP and now we can say that we created an instance and successfully created a database that people could access from anywhere.