Extract

The original data source for both the GPU and CPU prices were CSV files from Kaggle.

Transform

GPU: The GPU data was downloaded as a CSV file and was imported into a Pandas data frame. A number of columns that had unusable data were removed to make the data frame more usable. The data will be organized by price so the price column was reformatted to remove the dollar sign and converted into an integer. In order to match the CPU data to merge on later, the year month day date had to be converted into quarterly dates in a specific format. Once the group decided that we wouldn’t be using any information other than quarterly date and price I removed all other columns, and then grouped the data by quarterly date, averaging the price. Lastly, the data frame was exported as a CSV so that it could be merged with the other tables.

CPU: The CPU data was imported into a Pandas dataframe. The recommended price column was refined to drop rows with no data and with price ranges since we were only looking for data with one starting price. This column was then modified to become a float and to drop the dollar sign and commas within the numbers. Finally the data frame was narrowed down to only include types of data that we found most useful for the future analysis. These were Collection Name, Vertical Segment, Processor Number, Launch Date, Recommended Price, Number of Cores, Number of Threads, Max Memory Size, and max memory bandwidth.

FINAL: In order to create the final table we would be uploading, a second dataframe was created from each dataframe (Cryptocurrency, CPU, and GPU) that grouped the average price by release quarter. These will then be merged to create the final table.

Load: We used MySQL as our database to store the dataframes from the cpu, gpu, Bitcoin, Etherium, Litecoin, and Ripple dataframes. Different tables were created for each dataframe because we thought it would be more useful to keep the information separate. The data frames were converted into csv files and transported into another jupyter notebook that contained all the loading code. The python notebook was connected to a MySQL database. After creating the MySQL tables the dataframes were uploaded into respective tables. The index from pandas was transferred into a useless column in MySQL so we deleted the column.