**ETL Project: Group 4**

**Carfax**

**Team Members:** David Obembe, John Dilligard and Christion Lankford

1. **Introduction**

ETL is a type of data integration that refers to the three steps (extract, transform, load) used to blend data from multiple sources. An ETL process involves extracting data from a source system, converting (transforming) the data into a format that can be analysed and finally storing(loading) the data in a data warehouse or other system.

Objective:

Use newly learnt skills in extraction, transformation and loading of data to:

* Scrape information on pre-qualified “Jeep Cherokee” SUV’s for sale on the internet.
* Organise the data into comprehensive tables that can be accessed by a buyer.
* Arrange the tables into a PostgreSQL database or MongoDB.
* Automate the system so that the database is updated with the latest Jeeps anytime the project code is run.

1. **Data Source**

* Carfax.com
* Cargurus.com (This was our second website that we intended on scraping, but due to shortness of time we could not complete this process)

1. **Extraction Process**

Two files were created using Jupyter notebook —” CarFaxVehicleScrapper.ipynb”, and “ETL-project.ipynb”. The former was used for scraping Carfax data for a single car whereas the latter was used for scraping all the cars on the first page of the website.

We managed to make our code scrape data for one car, then once we got that to work we imported it as a function into the “ETL\_project.ipynb” file so that it could be used to loop through all the individual html pages of the vehicles in Carfax.

Each Vehicle that was scraped had information that was brought back in dictionary format. By the end of the process we had a list of dictionaries containing Jeep Cherokee Carfax information.

**4. Transformation Process**

The next stage of our project was to transform the data. We decided to organise the data in 5 tables — dealer info table, car description table, vehicle info table, drivetrain table and Price table. Using ‘for’ loops we created lists containing key information about each car that was initially stored in dictionaries and we went further to ‘zip’ these lists into dataframes using pandas ‘pd.DataFrame’. Every Dataframe had the Vin Number as the primary key.

**5. Loading into MongoDB**

The final stage of the Project was to insert all the data frames into a database. We chose MongoDB over PostgreSQL because it was flexible and schema-less. This meant that we could add new fields and modify our data in the future without affecting existing rows of information. MongoDB gave us the freedom to work with data with potential for rapid growth.

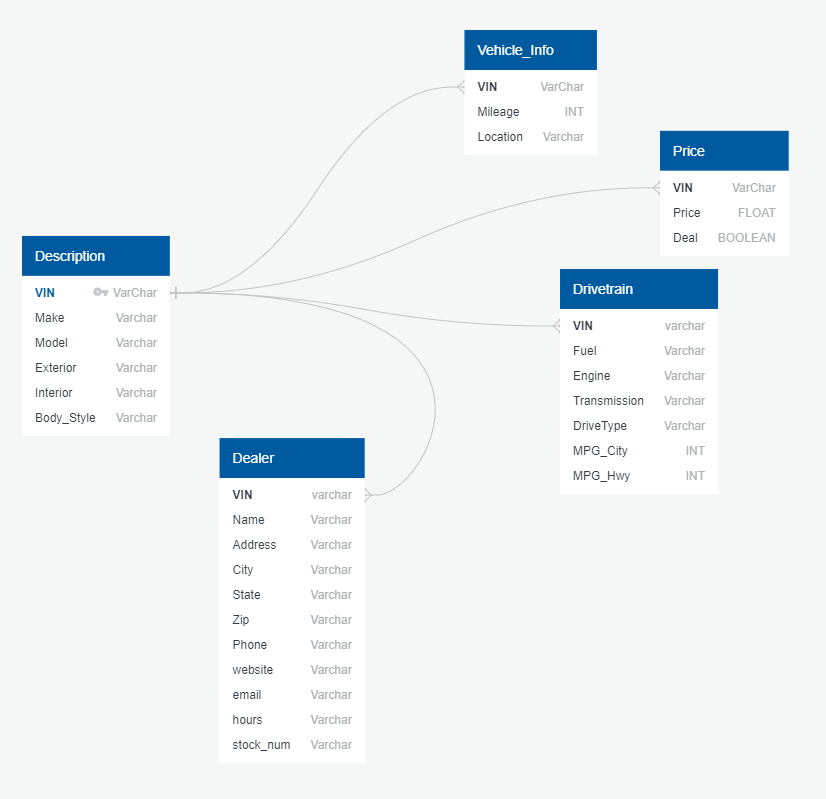
Our Final Database was called carfaxDB and it consisted of 5 collections that were imported via pymongo.

**6. Conclusion**

This was a mentally challenging project that expanded our understanding of how data pipelines are created, strengthened our skills in web-scraping and improved our knowledge of NoSQL databases. Our final product was a database that updates periodically and gives the user information of certified Jeep SUV’s that are available in the Texas area as well as their prices, interior design, model, engine, transmission, mileage etc.

Skills in teamwork, communication and time management were key in completing this Project as well as co-operation with our project supervisors.

**Appendix**

****

**Fig.1:** ERD diagram of database