ETL Project – Team JHD

Extract –

Web scraping the AMC and Cinemark websites for movies now playing. We want to compare the larger cinemas to see if there is overlap with newly released movies or exclusive movie releases between them.

Initially, we attempted to include Regal cinemas in our dataset as they are comparable in size to AMC but pulling images was unavailable to us. Regal was using a script to query images from a database, unavailable to web scraping.

Transform –

With each site, advance tickets sales are also promoted with now playing features. We had to design a way to remove these movie titles while scraping to eliminate the information up front.

AMC was scraped and loaded to a database using SQLite.

Cinemark was scraped using python, creating a dataframe prior to loading to a database.

Load –

SQLite was our method of creating a database for the information.

We joined the two tables using an outer join. The results will show the overlapping feature movies showing at both cinema chains will contain