# **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 for web scraping.

Cinemark was the next chain available but with about half the number of screens compared to AMC.

## Transform

With each site, advanced tickets sales are promoted along Now Playing feature films. 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 table. The table was created upfront and as the data was scraped, it was then loaded to the table.

Cinemark was scraped and each element (Title and Poster) was added to its own list. A dataframe was created, cleaned up, then a table was created and the dataframe contents loaded into the table.

Two different and effective ways to create the two tables: AMC and Cinemark.

## Load

We joined the two tables using a left/union all function to mimic the outer join function. A view was created (complete\_movie\_list) to display the contents. There were 15 new release movies consistent between the two theatre chains with Cinemark having 59 other films playing throughout their locations. AMC had 8 films currently being showed only in their theatres.

Technology

BeautifulSoup for web scraping, SQLite3 for table creation, Python/pandas for data manipulation.