ETL Project Report

Kim, Hunter, parmis, bita

# Extract

For this project, we decided to utilize open source movie data from Kaggle datasets. The main datasets we chose were:

* [The Movie Dataset](https://www.kaggle.com/rounakbanik/the-movies-dataset/downloads/the-movies-dataset.zip/7) – “Metadata on over 45,000 movies. 26 million ratings from over 270,000 users”
  + Data was downloaded as CSV files, with certain entries being JSON strings (lists of dictionaries). CSV files used were movie\_metadata.csv and keywords.csv.
  + All csv files included a movie id column that we used as our index value
* [The Academy Awards 1927-2015](https://www.kaggle.com/theacademy/academy-awards/downloads/academy-awards.zip/1) – “What actors and films have received the most Oscars?”
  + Data was downloaded as a single CSV file

# Transform

We primarily used PANDAS to clean and transform our datasets. Steps include:

* Loaded movie metadata into pandas dataframe;
* Dropped id columns that were loaded incorrectly (had a date string in the id column);
* Added id to AA table
* Jsonifyied AA table
* Merged keyword, meta data table

# Load

the final database, tables/collections, and why this was chosen as well as your final data schema.

* Monogdb

## Future WOrk

Create fun game of using keywords to guess movie title