**EMSE 6992 (DBMS for Analytics) Final Project**

Professor: Joel Klein

**Project:**

For this project you’ll be taking a dataset, loading it into a database, and then leveraging the database to develop some analytical insights. The datasets cover a variety of topics and formats. Your first task will be to determine which database system you think will work best with your selected dataset. At that point you will need to write a script, or leverage tools for the database, to load the dataset. Once the information has been loaded, you will need to query the database to create your insights.

**Datasets:**

* **Json:**
  + Amazon Reviews (Kindle Store): <https://www.kaggle.com/bharadwaj6/kindle-reviews/data>
  + Zomato Restaurants: <https://www.kaggle.com/shrutimehta/zomato-restaurants-data>
  + TV Shows: [https://github.com/jdorfman/awesome-json-datasets#tv-shows](https://github.com/jdorfman/awesome-json-datasets%23tv-shows)
  + [Jdorfman Json Datasets](https://json-datasets.zeef.com/jdorfman):
    - Reddit
    - GitHub API
  + Questions of Love: <https://www.kaggle.com/bobazooba/thousands-of-questions-about-love>
  + Yelp: <https://www.kaggle.com/yelp-dataset/yelp-dataset>
* **Other** (These datasets may require additional work):
  + Enron Emails: <https://www.cs.cmu.edu/~./enron/>
  + Million Song Database: <https://labrosa.ee.columbia.edu/millionsong/>

**Databases:**

* MySQL
* Mongo
* Dynamo
* Arango
* RDS (Amazon’s service for MySQL/SQL)