**EMSE 6586 (DBMS for Analytics) Final Project**

***Deliverable 1***

Professor: Joel Klein

**Objectives:**

The goal of the final project revolves around standing up a Document and SQL database, loading in a selected dataset, and then using the databases to analyze the data. To accomplish this task, all students can work independently or in groups of two.

To make this project more manageable, it will be broken into tasks and deliverables. The initial deliverable will replace the class midterm and will focus on setting up the project resources. For this deliverable you (and your team) will be required to send information on the selection of your dataset. Once a dataset is selected, you will be expected to setup your databases and being loading your data.

ETL-Extract-Transform-Load

**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/>
  + *Students may use other datasets with instructor approval*

**Databases:**

* SQL
  + MySQL
  + RDS (Amazon’s service for MySQL/SQL)
* Document
  + Mongo

**Tasks:**

1. Team and data selection (10pts):
   1. **Due: 3/20/2019**
   2. Email Submission detailing your team and dataset selected for the final project.
2. Standing up and preliminary population of databases for final project (10pts).
   1. **Due: 4/2/2019**
   2. Email submission with screenshots of collections and tables from you team’s databases.