**EMSE 6586 & CSCI 6907**

**DBMS for Data Analytics: Final Project**

Professor: Joel Klein

**Project:**

The final project for the class focuses on the intersection of data storage and analysis and is broken into four deliverables. You (and one teammate if you choose) will need to select from the provided project ideas listed below, standup a GitHub repository for collaborating/tracking progress, and put together a presentation to detail your findings. How each of these deliverables is assessed will be provided as each deliverable date gets closer.

**Project Ideas:**

|  |  |  |
| --- | --- | --- |
| Project Type | Description | Sample Datasets/Databases |
| Database Translation | Restructure a flattened and/or denormalized dataset so it can be loaded into a SQL or Graph database. Identify and demonstrate a form of analysis that is made easier by restructuring the data.  You are welcome to choose from the datasets provided on the right or propose your own dataset. However, any proposed datasets will require my approval. | * 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> * [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> |
| Additional Twitter Database | Using the twitter data from the class, standup a new database and detail how its functionality changes the way the twitter data can be used. | * S3 + Spark * AWS Neptune * Azure Cosmos (Graph API) * Postgres (with JSON fields) * Etc. |
| Twitter Analysis | Leveraging one or more of the databases provided throughout the class, analyze the elon-centric twitter dataset. |  |
| Custom Project | If there is an existing dataset or database that you would like to explore further, you can propose a custom project. However, these projects will be more heavily scrutinized before I will sign off on them to ensure they are feasible and a good fit for this course. |  |

**Deliverables:**

1. **Project/Team/Research (Due: 3/3)** – Identify your team, GitHub accounts, and project/approach along with any materials you have researched that will be the foundation for your project or approach.
2. **GitHub Repo (Due: 3/17)** – Create a GitHub repo within the [GWU-DBMS-For-Analytics](https://github.com/GWU-DBMS-For-Analytics) organization. The repo should be initialized with a couple of commits laying the foundation for your project.
3. **Implementation Finished (Due: 4/14)** – Deliver a working implementation of your code to ensure that you are on track for the final presentation. At this point your repo should provide information on where to find your data (if not provided in the repo), any ingestion scripts, any analysis scripts/notebooks, etc.
4. **Final Presentation (Due: 4/20)** – Present your results, detailing what your project is, how you implemented it, and a demonstration of your code and/or findings.