**Scope**: Migrating the 10K filings to MongoDB and conducting queries and analysis.

**Dataset**: Data has been collected for Dow Jones 30 companies and is available in a shared online folder (see instructions below). The files of interest for this mini-project are the 10K filings. This type of filing typically includes several parts (Item 1, Item 1A…) and sub-parts (company background, business strategy, products…). It also includes several Financial Statements sections. Your group will handle 3-4 companies.

The data in the 10K needs to be supplemented with additional firm level data (e.g., SEC identifier, company address) that you choose.

1. **Data Modeling**: Explain the pseudo data model that you used to store the data in MongoDB. In your answer make sure you also address the following

- What databases, collections, documents will you create?

- What constraints have you used, what indexes...?

- What format will you convert the data into?

1. **Data Migration**: Explain in detail your process of migrating the data from its current format (be it html or txt) to MongoDB. The numeric data in the Financial Statements sections need to be carefully migrated. Include a discussion of the intermediary steps (i.e, how to break down the full 10K filing into its parts and sub-parts). Implement your migration procedure into MongoDB for all the company datasets in your group. You will have to supplement the data in the 10K filing.

- Provide the needed details about the process as well as the implementation.

- Include snapshots when needed

1. **Company health**: Derive a metric for evaluating the company health using the Financial Statements sections.

- Explain the rationale you use.

- Write the needed queries.

- Summarize and present your findings.

1. **Acquisitions**: These events are typically reported in the 10K. Devise a way to identify the number of companies acquired by the mother company and the dollar value of these acquisitions.

- Explain the rationale you use for identifying the companies and the transaction values.

- How would you automate this process?

- What were the difficulties you encountered (which might result in inaccurate derivations of the values)?

Shared Drive Access:

Instructions on how to access the shared drive is available at <http://technology.gsu.edu/2014/02/10/access-department-shared-drive-windows/> or <http://technology.gsu.edu/2014/03/10/access-department-drive-mac/> for connectivity

Access using gsuad\CampusID (in lieu of step 2B)

The path to the Secure share access: [\\fsurprdext01.dmd.gsuad.gsu.edu\public](file:///\\fsurprdext01.dmd.gsuad.gsu.edu\public) (look for folder SECdata\**Institute**). Once the drives are connected as network drives, you can transfer data to or from just like a local data drive.

Additional data:

This SEC page allows downloading a wealth of company-related data. Note however that you are not required to use this data <http://www.sec.gov/dera/data/financial-statement-data-sets.html>

Notes

- The data was collected for 10K (with its htm, txt and xml formats) and 8K (with all its formats).

- Firms file amendments to their 10K and 8K. These are known as 10K/A or 8K/A. These files have also been collected.

- The data spans years 2010 to 2015.