IRisk Lab Data Discovery and Consolidation - Task 2
Week 1 Report
Aug. 29
Rohit Valmeekam

What was accomplished

1. Created rough guidelines for the creation of the database system. We decided to group by Insurance category, as the data that the Task 1 team has gathered is grouped by Insurance category.
2. I initiated the thought process of how our database system would end up as, and I decided on the creation of a one-to-many relationship between “Insurance Category” and “Dataset Table” database tables to thoroughly display all the datasets in an organized manner. This is due to the nature of the datasets that were acquired by Team 1 and them being grouped by different insurance categories.
3. Started thinking about what technologies to use and we ended up deciding on MongoDB to create our database system due to the fact that it is faster for large databases, whereas MySQL would be slower in comparison.
    a. Still in the process of deciding whether to utilize a Relational Database Management System or a Document Database.
        i. Document Database is more horizontally scalable however if the data conforms to a rigid structure, we may have to utilize a Relational Database System.
4. Started reading a textbook on MongoDB to familiarize myself with the technology.
5. Sent email to Eli O'Donohue asking for more information on the steps necessary to create a Relational Database Management System, if we were to utilize one over a Document Database.

To-Do List

1. Learn Database Systems Concepts
    a. Chapter 2 to Chapter 5 of PDF (Database System Concepts, Abraham Silberschatz)
    b. MongoDB: https://www.mongodb.com/docs/manual/introduction/
2. Learn how to utilize the NCSA Deep Learning Cluster (How to collaboratively build database system on our HAL clusters)
3. Data pre processing (Remove some invalid data; Data cleaning)
    a. Communicate with Task 1 about starting the pre-processing stage of the data as they are collecting datasets so that we can start creating our database system sooner
4. Integrate data in different file types
    a. The data collected by Task 1 has lots of different file types, so being able to integrate data in different file types is a challenge we have to start exploring
5. Write scripts to import data to our database system (data pipeline creation)