Project Plan - Buchanan Summer 2016

Alec Buchanan

June 10, 2016

Hours: 7:30 am to 5:00pm

## Description

Both MongoDB and Neo4j are NoSQL databases (Not Only SQL). MongoDB is a powerful document database, while Neo4j is a powerful graph database. Using these two technologies in unison, we are able to see relationships between data that we would have never seen before. Right now, the technology is not very well known and people do not understand what it is capable of. This is why David Meza and I have set out to look for ways to apply this new technology. We were looking at using this system to help project managers comb through thousands of incident reports. The project managers would also be pointed to relevant documents that they would not have thought of checking. Due to inconsistencies and unorganized lesson learned data, we are looking at applying this technology to research experiment results from the International Space Station (ISS).

## Plan of Attack/Objective

I am starting with a little help. The previous intern left some documentation on how to setup both MongoDB and Neo4j. He also left information on how to connect the two and a python parser. The python parser was designed to break up certain lesson learned files, but I may be able to repurpose parts of it to scrape and parse ISS research information that can be found on [NASA’s public website](http://www.nasa.gov/mission_pages/station/research/experiments/experiments_hardware.html).

Before we are able to showcase any of this new technology we need to have data to put into the database. I will be taking raw data from NASA’s website about research experiments performed on the ISS between 2000 and 2016. I will write a python script that will scrape the webpages and format the information into a JSON file. The program will then import the information from the JSON file into MongoDB. Since we have MongoDB and Neo4j connected, MongoDB will automatically push information to Neo4j. If we have time, we will setup Linkurious to make the information more readable for end users.

### Steps:

1. Create python script to scrape and parse the raw data from NASA’s website.
2. Setup MongoDB and connect Neo4j. Do the same thing on my desktop for testing. A graph model also needs to be thought out for Neo4j.
3. Find a way to generate relationships between experiments
4. etup Linkurious so people can easily interpret and visualize the relationships.

*Timeline is somewhat dependent on when I will be granted escalated privileges.*

## Appendix

**Document Database:** A document-oriented database, or document store, is a computer program designed for storing, retrieving, and managing document-oriented information, also known as semi-structured data.

**Graph DataBase:** In computing, a graph database is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data. A key concept of the system is the graph (or edge or relationship), which directly relates data items in the store.

**Graph Model:** A graphical model or probabilistic graphical model (PGM) is a probabilistic model for which a graph expresses the conditional dependence structure between random variables. They are commonly used in probability theory, statistics—particularly Bayesian statistics—and machine learning.

**JSON:** JSON is an open-standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs.

**Linkurious:** Linkurious allows end users to visualize data in graphical databases. Linurious is most commonly linked up with Neo4j.

**MongoDB:** MongoDB is a free and open-source cross-platform document-oriented database. Classified as a NoSQL database, MongoDB avoids the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster.

**Neo4j:** Neo4j is a graph database management system developed by Neo Technology, Inc. Described by its developers as an ACID-compliant transactional database with native graph storage and processing, Neo4j is the most popular graph database according to db-engines.com.

**NOSQL:** A NoSQL (originally referring to "non SQL" or "non relational")[1] database provides a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases.

**Scraping:** Web scraping (web harvesting or web data extraction) is a computer software technique of extracting information from websites. This is accomplished by either directly implementing the Hypertext Transfer Protocol (on which the Web is based), or embedding a web browser.