Big Data Architecture - SEIS736

Project Proposal

Wade Lykkehoy

Student ID: 101217575

lykk3260@stthomas.edu

**Data Sources**

White House visitor logs from the Obama administration. Approximately 6 million records; 1.1 Gig of data.

My contacts and LinkedIn connections; 462 records.

**Analysis Approach**

1. Data profiling and cleansing

For the data fields I will be using, determine the data types (string, numeric, categorical, etc) and profile the data for range of values, blanks, invalid values. Based on this information, perform any required data cleansing.

1. Generate descriptive statistics

I envision generating a variety of descriptive statistics based on the data. Following are some of the items I envision; I expect once I get more into the data, additional ideas will pop up.

General:

Top reasons for visiting.

Top individuals visited.

Percent of visitors on a tour, visiting POTUS, visiting FLOTUS. Plus others depending on what is learned when analyzing reason for visiting.

Visitors on tours:

Number of tour visitors by year and month.

Seasonality; was there a time of year when there were more tour visitors; e.g. spring, summer, fall, winter.

Based on visitor name, does there appear to be repeat tour visitors.

POTUS visitors:

Identify repeat visitors.

FLOTUS visitors:

Identify repeat visitors.

Join to my contacts and LinkedIn connections:

Did anyone I know appear to have visited the White House.