Building Better Nonprofits

W205 - Milestone 2
Ashley, Ganesh, Matthew, Razib

Problem & Motivation: With the growing number of nonprofit organizations(NGO), data science can help these organizations steer their limited resources towards the right target for maximum impact, as well as help policy decision makers on defining the scope of these nonprofit.

Previous Research: Organizations like Guidestar [www.guidestar.com] do provide NGOs access to resources based on NGO IRS information and information that NGOs input into Guidestar; however, it does not combine the IRS data with other geographic or locational information and social media data that may prove to be a factor in NGO success.

Data Sources: To build off of previous work that has been done, we propose to include additional datasets into our analysis and schema to help identify external possible geographic factors that may contribute to NGO success. Our data sources together incorporate elements of the 3 V's:

- Volume - There are multiple large datasets including tax data for over 1 million NGOs.
- Velocity - The IRS data is updated monthly and the population and GDP data are updated yearly or quarterly. Additionally, the social media data like twitter may provide a streaming near-real time velocity aspect..
- Variety - Each dataset has its own unique features and different updating properties which provide variety in the sourcing and handling of each dataset.

The primary dataset is focused on the yearly tax returns filed by the nonprofits. AWS hosts a master dataset which includes basic information for all the nonprofits. [https://aws.amazon.com/public-data-sets/irs-990/]. This information will be used to obtain individual tax returns (IRS 990 tax forms) which are also available through AWS. Geographic and social media datasets will be joined to gain more insight about the performance of the organizations.

In order to pull in more geographic information, we are proposing using two additional datasets related to population and economic factors of different states and metropolitan areas. A third dataset (Twitter) will provide the social media aspect to nonprofits.

- The US Department of Commerce Bureau of Economic Analysis provides public access to a variety of datasets, we have pulled the datasets on gross domestic product (GDP) in the units of current dollars for both the state level and the metropolitan areas. [http://www.bea.gov/regional/]
- The US Census Bureau provides public access to a variety of population related

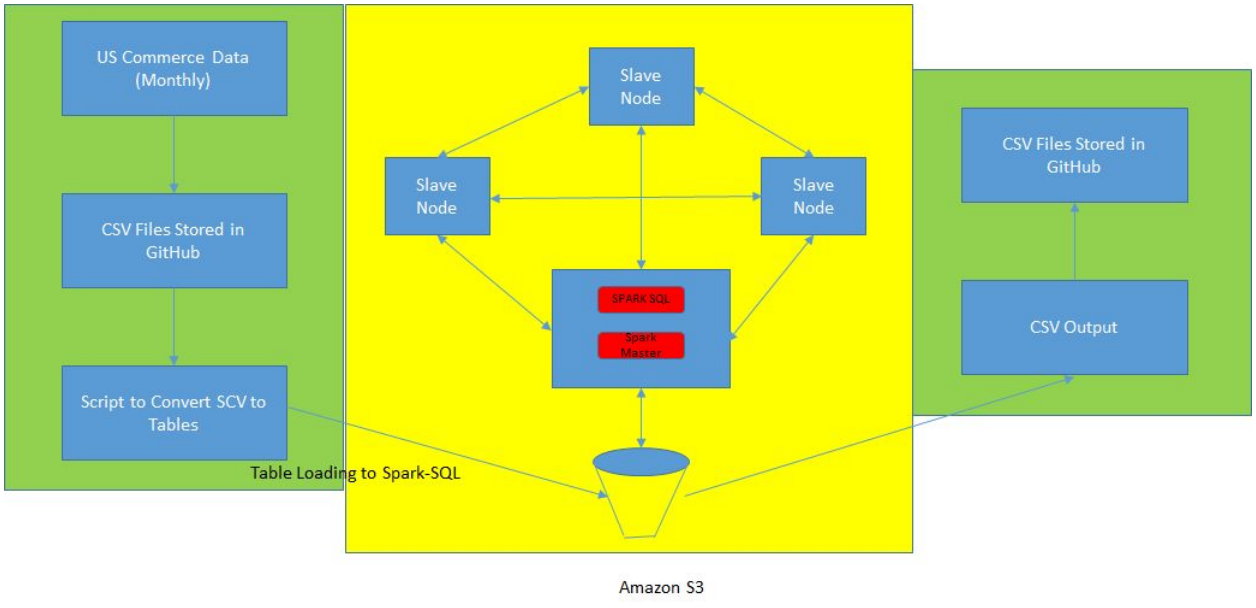
datasets. Similarly to the GDP datasets, we have downloaded and started to explore both the State level data and the Metropolitan level data. [https://www.census.gov/popest/data/datasets.html]

- Twitter provides two APIs for developers to access data, a REST and Streaming API. The streaming API will be used to analyze and store tweets in real-time that match our nonprofit search criteria. The REST API will be used to supplement data pulled from the Twitter stream.

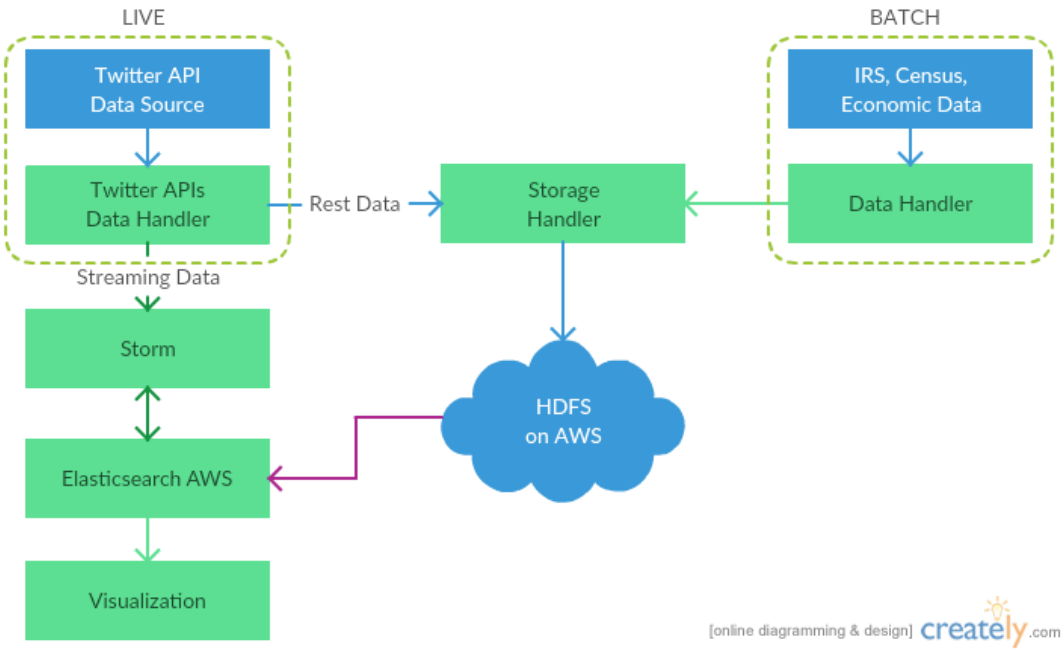
Schema: The primary dataset on NGO tax information has an associated schema already, but in order to incorporate our other datasets, we will redesign the schema to include the appropriate information and key entities and relationships from all of the datasets.

Storage & Architecture: The semi-static master datasets will be stored on Github for easy access in one place. However, the actual work will be done on AWS and in that environment the data and associated schema will be done using Spark-SQL for semi-static data based US commerce website. The semi-static data from in the website is uploaded on a monthly basis. Non-static data like streaming data from social media will be handled using Apache Storm.

Spark SQL will be used to create tables and draw inferences for semi-static data. The following figure shows the dataflow/architecture of the project for semi-static data.



Below is a high-level view of the entire system architecture, which includes the streaming data twitter data.



Outcomes:

Our goal is to find out

- Which type of nonprofits are most successful and which are most unsuccessful
- Is there a correlation between nonprofits success/impact to
    - The number of employees
    - Employee compensation
    - Employee-volunteer ratio
    - Social media presence

We will also include geographic and social media information to

- Determine the population preference for charities within a locality
- Which NGOs are effective in their fund utilization
- How successful are similar nonprofits at different locations
- Sentiment analysis about nonprofits