Building Better Nonprofits

W205 - Milestone 3
Ashley, Ganesh, Matthew, Razib

Problem & Motivation Summary: The problem and motivation has remained focused on building better nonprofits and links to geographic locations. 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 instead of having to employ data scientists in each of these organizations. In addition this work can help new NGOs make decisions on the best locations to start up.

Data Acquisition & Organization: The majority of the time spent on this project so far has been focused on the organization and implementation of the schema and testing the architecture. The details provided below have been tested on small subsets of the data and the entire data, since the scripts are independent of the data size.

The IRS 990 dataset contains a index.json file which has 3.5 million entries and an abundance of fields. We have chosen to filter the data to include the fields listed below that are most relevant for the project:

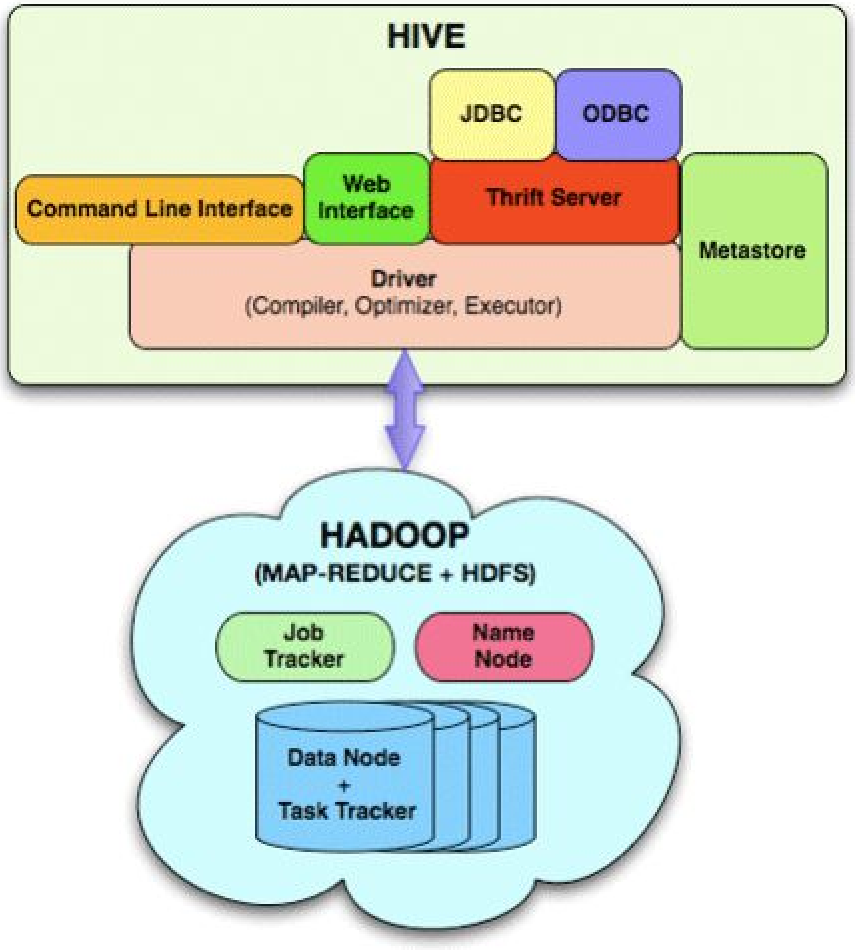
- EIN of a non profit,
- Name of that organization,
- Tax period for that tax return,
- the form type of the Tax return (the tax return form type changes based on the annual gross receipt of the organization)
- URL for the electronic copy of the tax return (these are xml files)

We have created a python script to convert this index.json file to a csv and divided it into 35 files (620 MB) so it can be loaded into Github. A python script has also been created to read all the tags and corresponding text of the xml files.

The geographic population and GDP data are fairly small (less than 1G) flat, structured, csv files. This data is stored on github and can be pulled into the AWS environment through a git pull and then added into the schema.

Since some of this data does get updated on a regular basis, future ingestion of the data can easily be done by re-running the scripts. Additionally, fields can be easily added or changed by modifying the same scripts to explore different questions.

Architecture Progress: The high-level architecture and the lower level components of the architecture are shown below



Raw data generated from the CSV files using PERL scripts are loaded into HDFS and raw schemas generated from the headers of the CSV files are loaded into Hive. Data ingestion to HDFS-Hive will happen on a monthly basis whenever IRS data is updated with new information. Currently the plan is to have a manual update (as an improvement, the update can be automated in the future).

Data clean up involves removing incomplete data, any outliers/irrelevant data from the tables uploaded into hive. Secondary schemas have been defined to create new tables using only relevant fields for the project. Final visualization of data is done using Tableau which is connected to the hiveserver2. .

Analysis: In order to provide useful analysis with this data, we need to define a measure or proxy for “success” of an NGO. For the purposes of this project we are using the definition of overall growth or money received is within +/- 5 percent of the previous years. This definition of success should avoid biases for larger NGOs and provide a useful measure for comparison. This will then be combined with the geographic and location data for possible identification of patterns that contribute to success.

Additionally, the twitter data may be able to provide the sentiment component of the data. This would come in the form of words associated with the NGO in tweets and the frequency of words. Additional analysis can then take the top words and provide insight into positive or negative sentiments.

The end goal and resulting visualization or user interface/interaction will be done through Tableau and creating an interactive dashboard for the user to explore the results.

Adaptations to Original Strategy: We still have the plan to include some twitter data but are waiting to complete exercise 2 in order to implement this fully. However, time may not allow for this addition. The streaming twitter data may end up being too much to add into this project because of the additional text analysis required. The twitter data could provide additional insights but due to the need to finish exercise 2 in order to really learn how to use and leverage the streaming data, the time remaining in class may not be enough time to incorporate this into the final project.