Extracting Insights from FBO.Gov data - Part 1

Extracting Insights from FBO.Gov data - Part 1

Earlier this year, Sunlight foundation filed a lawsuit under the Freedom of Information Act. The lawsuit requested solication and award notices from FBO.gov. In November, Sunlight received over a decade's worth of information and posted the information on-line for public downloading. I want to say a big thanks to Ginger McCall and Kaitlin Devine for the work that went into making this data available.

From the Sunlight page linked above:
"The notices are broken out into two parts. One file has records with the unique id for each notice, and almost all of the related data fields. The other file with the **_desc** suffix has the unique id for each notice and a longer prose field that includes the description of the notice. Additionally, the numbers at the end of each file name indicate the year range." So what we have are two files for each year. One file has the metadata about the solicitation and award and the other related file has the verbage about the notice.

The metadata consists of the following rows:

- Posted Date
- Class Code
- Office Address Text
- Agency Name
- POC Name
- POC Text
- Solicitation Number
- Response Deadline
- Archive Date
- Additional Info Link
- POC Email
- Set Aside
- Notice Type
- Contract Award Number
- Contract Award Amount
- Set Aside
- Contract Line Item Number
- Contract Award Date
- Awardee
- Contractor Awarded DUNS
- JIA Statutory Authority
- JA Mod Number

- Fair Opportunity JA
- Delivery Order Number
- Notice ID

The metadata file is a .CSV approximately 250-300MB in size. It is a relatively large file that isn't easily manipulated in Excel. Additionally, it has embedded quotes, commas, html code and multiple lines with returns in the field. For file manipulation, I decided to use Vertascale. Vertascale is designed to provide real-time insights for data professionals on data stored in S3, Hadoop or on your desktop. It can easily handle text, json, zip, and a few other formats.

The decision to use Vertascale was driven by the size of the file, the ability for it to read the file correctly (i.e. parse the file correctly) and the ability to transform the data out into a format that I could use with other tools. Let's run through how Vertascale can do this quickly and easily:

A) Opening & converting raw data into a more useful format

Step 1: Open the Vertascale application and navigate to the file that is on my local machine. In this case, we will use:

Step 2: Next, use Vertascale's built in parsing tools to convert the file into a more user friendly view:

Step 3: Export the columns into a file that we can use later.

B) Getting a feel for the data

Now that the data is loaded in Vertascale (which can be an overwhelming task at times), we can explore and analyze the data. We will look at the Browse and the Count feature of Vertascale to understand the data.

1) Browse
I used Vertascale's built-in data browser to browse through the dataset to see what data was available. With the slider and next 50 options, I could quickly go to any spot in the file and see what the data was available.

2) Count
I wanted to look at the "Set aside" column to see what the distribution of set aside values where in the file. Using the "Count Distinct" function, I was able to see how those values were distributed across the entire file. In this case, we can see

Next Steps

At this point, we have retrieved the FBO datafiles, used Vertascale to get a feel for what is in the files, extracted columns out to do some initial analysis. In the next post, we'll dive in a little more to look at some ways of analyzing and visualizing this data.