Project Note for GoRIM application

Yina Xu, Ari. 21, 2019

Requirements of the application

- build a web based application
- importing/uploading CSV files
- data manipulation such as joining multiple tables
- data analysis such as find correlations between features
- visualize the analytic results

Weakness of Watson Analytics:
- doesn't support 3+ table joins (only support 2)

1. Common issues of open source BI tools

After investigating 8+ open source tools, the following is the common issues from them:
pros:

- great user interfaces
- visualize data with multiple rows
- look into raw table data on web
- writing and running SQL queries (few support)

cons:

- don't support importing CSV files
- can't manipulate data such as table joins
- don't support data analysis (most of them focus on the display data from database)
- poor documentation (few)

Since open source tools don't support data manipulation as well as strong analytic abilities which are key in (for building) GoRIM, we need to build our own application.

** The following is the list of open source BI tools that were investigated based on Github BI Ranking **
The numbers in bracket are stars in their Github repositories.

1. metabase(13k): has a very friendly UI, but only support uploading/displaying tables, doesn't support table join & data analysis
2. cBoard (1.7k): doesn't support import CSV & data analysis

3. just-dashboard(1.3k): focus on data visualization
4. mining(0.9k): it only supports Windows based development but doesn't mention in the document (until I executed on mac and found the exact error line of code)
5. dbt(0.6k): doesn't have a clear document, failed to execute
6. tabix(0.6k): doesn't support data analysis, more focuses on SQL manipulation & data visualization
7. Seal(0.6k): not a web application, Windows based application
8. squealy(0.6k): doesn't support data analysis, API doc UI

2. Recommendation in Data Analytic Language

Python:
pros:

- one of the most popular data science language as well as server side language
- have many open source data science tools & frameworks
- have large community that can easily find learning/debugging source on the website
- have relatively low learning curve compare to other server side languages such as Java, C#

cons:
not found yet :P

R:
pros:

- widely used for data science and statistical computing
- easy to learn and apply
- have a large community and resources on the internet

cons:

- not a server side language
- importing/exporting data to the server side require extra effort

Based on the survey, in 2018, the most popular language was Python with 65.6% people used it as data analytic tool, with the 11% growth from 2017. There are 45.8% people showed using R, dropped 14% from 2017. This shows Python is the dominant language in data analytics and supporting from many developers. Thus, Python is highly recommended compared to R language.

3. Recommendation in Data Analytic Tool
Recommended:
Pandas (18k):

- an open source data analytic tools
- one of the most popular data analytic Python tools
- focuses on data analytic and provides easy and intuitive APIs to use
- low learning curve for developers with limited machine learning knowledges
- project is still active and maintained and updated by authors frequently
- support import/export CSV files (verified: pandas.read_csv)
- support 3+ table joins (verified: pandas.merge) and many other data manipulation operations
- support finding correlation between columns (verified: pandas.DataFrame.corr)

** tested with Helloword dataset, an example of finding correlations between columns **

```
                          Advertisement Program  ...  Annual Monitored Indicators
Advertisement Program                  1.000000  ...                    -0.934747
Number Reached by Facebook             0.897099  ...                    -0.704183
Annual Monitored Indicators           -0.934747  ...                     1.000000

[3 rows x 3 columns]
```

others: (all support data manipulation and data analytics)

- Tensorflow (120k): The most popular Python based tool, focusing on machine learning. Thus, it has relatively high learning curve compared to Pandas since it requires ML related knowledge. The repository is still active and maintained by Google.
- Pythorch(24k): A deep learning platform for researchers maintained by Facebook. Since our dataset won't be too large, we don't need deep learning tool which increase the learning difficulties.
- Numpy(10k): It is more popular to be used in calculating scientific/mathenmatic computing.
- JupyterLab(7k): Jupyter provides a web platform for data analysis who can write python code and see data visualization directly in the browser. It is not UI friendly (code only) and not easy to extend the platform.

Although there are more popular Python based frameworks, Panda is highly recommended after comparing props/cons and it meets all our requirements.

4. Recommendation in Data Visualization Tool
The most of data visualization tools support displaying chart with multiple rows, provide many options (pie, line etc.). Thus some other aspects are considered in this case.

Recommended:

- Chart.js (43k): An open source chart library with friendly user interface, easy to start, can write less code and do more stuff.
- Highcharts: Not an open source project, but offers limited free use of Canada Map.

others:

- D3 (82k): cumbersome and requires many code to implement one chart.
- Echarts(32k): high learning curve with too many parameters need to be used, poor English document
- Chartist-js(11k): no longer maintained, remains many Github open issues and last update was 5 months ago
- dash(8k): poor document in API guide (one short web page was all I found)

Overall, Chart.js is strongly recommended because of its strong document, popularity and large community compared to the other options.