Our goal is to develop an application that takes contract opportunity and job data from multiple sources and analyzes them to give Raytheon indicators of what contracts could be advantageous to the company.

**PROJECT SUMMARY**
- Our goal is to develop a dashboard that makes data efficiently visible and provides insights.
- Using Tableau to create a dashboard due to its efficiency and ease of use.
- Data structure can be changed over time.
- MongoDB Compass: a database is an organized collection of structured information.
- Made available to users for easy visualization and use.
- NLP uses research in linguistics to develop tools for topic modeling.
- Job opportunity displays areas that are in high demand.

**WEB SCRAPING**

**METHODS**
- Web scraping – utilized Beautiful Soup (Python)
- Automation – utilized Helium (Python)

**SOURCES**
- Multiple sources were used to diversify the data we collected. Both websites were free to use and scrape data from, practical and cost-efficient, while providing useful insights.
- Sam.gov
- Usajobs.gov

**RESULTS**
- Retrieved over 6 million contract opportunities.
- Contract opportunities are procurement notices from federal contracting offices.
- Anyone interested in doing business with the government can use this system to search opportunities.
- Successfully implemented the ETL (Extract, Transform, Load) process.
- Wanted both: Contract opportunity display the amount of budget invested in certain field, Job opportunity displays areas that are in high demand.
- Gathered over 13 thousand government job opportunities.

**DATABASE**

**GOALS**
- Get data that was extracted via Web Scaping and upload and store it in a MongoDB database for later usage and analysis.

**LIBRARIES AND INFRASTRUCTURE**
- MongoDB: a non-SQL database that can accept and save data.
- MongoDB Compass: an application that allows for direct access to a MongoDB database.
- Python Flask API: a website framework that can handle the functionality for saving and reading scraped data from the web.
- Dockers: a software that converts our code into a container (a package of our code and all its dependencies so the application can run quickly and reliably from one computer to another).
- Kubernetes: a library that acts like a manager to run and operate these containers.

**PROCEDURE**
- Load the data into a Mongo database that runs on Kubernetes container in Purdue’s Geddes server.
- Automatically upload data to the database using the MongoDB address.

**MongoDB ALTERNATIVE: APACHE DRILL**
- Explored as an alternative and backup to MongoDB in case infrastructure issues prevented the use of MongoDB.
- MongoDB data source can be connected to Apache Drill.
- Apache Drill has applications with MongoDB.
- Query data/data sets, configure storage, execute SQL queries.
**NLP**

**GOALS**
- Use natural language processing to extract data from the database to explore topic modeling
- Topic modeling – way of identifying patterns in text; topics are these patterns

**PROCESS**
- Retrieve data from the database in the form of text files
- Import packages (Python libraries with various functions)
  - NLTK, gensim, NumPy, pandas, matplotlib, PIL
- Clean data (stopwords, word_tokenize)
- Model topics using LDA (LdaModel function)
  - Latent Dirichlet Allocation (LDA) – Python function that organizes the text in a document by topic; it is an example of a topic model

**RESULTS**
- Our primary objective for this NLP code was to organize data for topic modeling
- This allowed us to identify multiple patterns in our data, group it properly and find trends within our data

**FIGURE 1**
- Figure 1 (bottom left) is a data visualization produced by querying the data in the database utilizing MongoDB Compass
- Shows the data sorted and analyzed by Department Agency

**FIGURE 2**
- Figure 2 (to the right) is a visual of topic modeling
- We aggregated data from the 2018 and 2021 National Defense Strategy and performed topic modeling using natural language processing (NLP)

**FIGURE 3 (ABOVE)**
- We used Tableau to create an interactive dashboard to easily view the analyzed data in a user-friendly manner
- The dashboard is simply a way to visualize interesting correlations with the data.
  - Figure 3 (below) shows various visualizations of the data ranging from job counts, Departments/Agencies vs. Sum of Awards, and dollar maps.

**CONCLUSION**

**STRENGTHS**
- We created a Web scraping process that was powerful and efficient allowing for capturing millions of jobs and contract opportunities.
- In-lab communication was great in terms of dividing up work based on each member’s strengths.

**CHALLENGES**
- Hosting the database on a server was a huge challenge and setback in terms of time and lack of knowledge and experience.
- Sub-team communications could be improved in the future, as one sub-team’s code can often be used in another sub-team’s work as well.
- Outside-of-lab communication can be improved in terms of checking in with each other on our progress.

Overall, this semester, we were able to successfully create the main Web scraping and Database infrastructure and created a solid base for further analysis and visualization as well.

**FUTURE PLANS**
- More data sources (ex. FRED St. Louis Fed)
- Continuous stream of data/automated pipeline
- Complete Flask database
- More descriptive analytics
- Even more detailed visualization tool with Tableau

**REASONING**
- Incorporating new data sources will provide more data
  - Leads to more descriptive analytics and better insights from the tool
- Automating the transfer of data will allow for constant pulling of data online and sending to database
  - Removes manual work
- Completing the Flask database gives a more desired alternative to the current MongoDB
- Since we have all the pieces, we simply need to put it together and create a final dashboard to display findings

**ACKNOWLEDGEMENTS**

We would like to thank Raytheon Technologies for the opportunity to take part in this project, and our Corporate Partners Mentor Mike Douglas for all his guidance and support throughout the development of our project. We would also like to thank the Purdue Data Mine Staff, Dr. Ward, Maggie Betz, and Kevin Amstutz.
<table>
<thead>
<tr>
<th>Job Name</th>
<th>Count</th>
<th>Award ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optometrist</td>
<td>84</td>
<td>420,837,6</td>
</tr>
<tr>
<td>Medical Support Assistant</td>
<td>28</td>
<td>14</td>
</tr>
<tr>
<td>Medical Record Technician</td>
<td>28</td>
<td>14</td>
</tr>
<tr>
<td>Dental Assistant</td>
<td>28</td>
<td>14</td>
</tr>
<tr>
<td>Clinical Psychologist</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Suicide Prevention Specialist</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Security Guard</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Pharmacy Technician</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Pharmacist</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Nurse (Community Health)</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Maintenance Worker</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>IT Specialist</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Health Technician (Mental)</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Health Technician (Cultural)</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Dental Assistant/Dental As</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Child and Youth Program As</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

**Agencies**
- Air Mobility Command
- Indian Health Service
- Veterans Health Administration

**Department/Ind. Agency**
- DEPT OF DEFENSE
- HOMELAND SECURITY, DEPARTMENT OF
- VETERANS AFFAIRS, DEPARTMENT OF

**Dollar Map All**

**Dollar Map AOL**

**USA Job Name Count**

**The Data Mine Corporate Partners Symposium 2022**