Department of Homeland Security
Solicitation Number # 70SBUR19Q00000066

Modernized DevSecOps and Analytics Services II

Provided to:

Hollie Walsh
Contract Specialist
Department of Homeland Security
70 Kimball Avenue
South Burlington VT 05403
Hollie.L.Walsh@uscis.dhs.gov
ContractingProposals@uscis.dhs.gov

Provided by:

Guidehouse LLP (formerly PricewaterhouseCoopers Public Sector LLP)
Shannon White, Partner
1800 Tysons Boulevard, 7th Floor
McLean, VA 22102-4257
Telephone: (571) 296-2571
swhite@guidehouse.com
www.guidehouse.com

Part 1: Factor 1 – UserStories.PDF
May 31, 2019

GSA Schedule No. 70: GS-35F-0263P
TIN: 82-4596065
DUNS: 079529872
CAGE Code: 783T6

2019-420

This proposal includes data that shall not be disclosed outside the Department of Homeland Security (DHS) and shall not be duplicated, used, or disclosed in-whole or in-part for any purpose other than to evaluate this proposal. If, however, a contract is awarded to this Contractor as a result of, or in connection with, the submission of this data, DHS shall have the right to duplicate, use, or disclose the data to the extent provided in the resulting contract. This restriction does not limit DHS’s right to use information contained in this data if it is obtained from another source without restriction. The data subject to this restriction are contained in all pages/sheets herein.

UserStories.PDF

1.1.1 Core user stories

User Story 1: As a product owner, I need to be able to review data and re-collect all the information used for the research.
Tasks:

- Research and identify best data sources
- Set up API to collect 10 years of data on company filings such as 10Ks from EDGAR Securities and Exchange Commission (SEC)
- Develop R scripts to conduct web-scraping of company websites for keywords and descriptions
- Develop R scripts to pull Wikipedia data for all companies available
- Develop R scripts to scrape 10 years of Yahoo Finance historical stock data for all companies available
- Develop scripts to automate all data collection and re-collection processes across parallel microservices

Acceptance criteria

- Given that I am a data scientist, when I select the export button, the raw data pulled from EDGAR SEC, Web, Yahoo Finance, and Wikipedia will export to a R or Python file

Story point 16

User Story 2: As a product owner, I need to be able to generate a list of 10 sectors and no more than 100 industries classifications group under the sectors with related companies weighting from collected data by running no more than two commands.
Tasks:

- Run data though Doc2Vec Natural Language Processor (NLP) model
- Design hierarchical clustering algorithm
- Cluster companies into 10 sectors and 100 industries
- Develop parameters to weight companies
- Build weighted bar chart widget

Acceptance Criteria:

- Given I am a user, when I navigate to the Alpha Taxonomy (AT) homepage, then I see all 10 sectors as individual tiles
- Given I am a user, when I select a sector, then I can see all industries under that sector, and when I click into each of the ten sectors, then I do not see more than 100 industries total
- Given I am a user, when I select any of the industries, then I can see all companies listed in that industry, and then each company is weighted based on 10 years of historical data

Story Points: 16

User Story 3: As a product owner, I need to be able to review classifications using an infinite scroll HTML web page. The web page shall include data visualization, analysis and insight of sectors and industries with recommended weighting of each company based on 10 years of historical data.

Tasks:

- Create and validate the UI wireframe
- Conduct A/B test UI testing on wireframe
- Leverage UI wireframe to develop front-end interphase
- Integrate interphase with solution
- Develop infinite scroll data table
- Develop main angular html template

Acceptance Criteria:

- Given that I am a user, when I navigate to the AT homepage, then the page is an infinite scroll HTML
- Given that I am a user, when I select a sector, then all information is available under all sector tiles, and then focus is on the selected sector information that was just made available
- Given that I am a user, when I select an industry, then all the industry specific information that is made available includes a list of companies related to that industry, and then all companies have a recommended weighting based on 10 years of historical data

Story Points 16

User Story 4: As a DevSecOps engineer, I need to be able to stand up entire infrastructure using 100% automated scripts with a minimum number of commands.

Tasks:

- Setup microservices, serverless functions, and cloud services/endpoints for solution to function
- Setup Terraform environment and scripts for all OpenShift containers and AWS Services
- Build out CI/CD (Continuous Integration/Continuous Delivery) pipeline
- Create tests to automatically verify build and troubleshoot for errors
- Consolidate deployment to single unified script

Acceptance Criteria:

- Given that I am a DevSecOps engineer, when I run the two commands described in the AT Notebook, then the entire infrastructure for the AT solution is stood up

Story Points 16

User story 5: As a data scientist, I need to be able to have an infinity scroll infographic to view the raw data.

Tasks:

- Leverage UI wireframe to design pages that have infinity scroll
- Write R Python script to pull all data
- Organize raw data so it is useful for the user
- Develop infinite scroll data table
- Develop main angular HTML template

Acceptance Criteria:

- Given that I am a data scientist, when I navigate to the AT homepage and select any button on the page, then all new content is displayed on the same page, and then I can see all information by scrolling
- Given that I am a data scientist, when select the export button on the data tables, then all raw data is exported in to R or Python

Story Points 8

User stories 6: As a data scientist, I need to be able to understand the classification and grouping of companies in sectors and industries.

Tasks:

- Define and develop Sector/Industry cluster mappings
- Develop code to scan and pull key indicating words
- Define each sector
- Define each industry
- Define typography to name clusters
- Develop UI wireframes that support clear classification / groupings

Acceptance Criteria:

- Given that I am a data scientist, when I am on the AT homepage, then I can see all 10 sectors tiles
- Given that I am a data scientist, when I select a sector, then I can see all industries and companies related to that sector
- Given that I am a data scientist, when I open the models.md posted in GitHub, then I see all clusters maps

Story Points 8

User story 7: As a data scientist, I need to be able to review parameters for the weighting of companies in each sector and industry.

Tasks:

- Collect data from over 10 years and analyze
- Define parameters for weighting companies
- Use historical stock price and market cap data to determine company vs. sector/industry weights

- Leverage cosine similarity to determine weights of industries based on closeness in the cluster map
- Design interface that will clearly illustrate company weights

Acceptance Criteria:

- Given that I am a data scientist, when I navigate to the data table, then I see all companies under the selected industry with a bar chat indicating their weight
- Given that I am a data scientist, when I open the models.md document posted in GitHub, then I can see the methodology for weighting companies

Story Points 8

1.1.2 Additional user stories

User story 8: As a Data Scientist, I want raw and abstracted code so I can easily use and edit it as needed.

Tasks:

- Provide cleanly commented R and Python code in GitHub
- Provide descriptions on the models used in a .md file

Acceptance Criteria:

- Given that I am a data scientist, when I open the R and Python code posted in GitHub, then I can access all raw and abstracted code
- Given that I am a data scientist, when I open the models.md file also posted in GitHub, then I can see all descriptions of the models and clusters leveraged in the AT solution

Story Points 3

User story 9: As a DevSecOps Engineer, I need a continuous integration pipeline to run automated tests so I can prevent faulty code from being promoted to the production environment.

Tasks:

- Build out CI/CD (Continuous Integration/Continuous Delivery) pipeline
- Create tests to automatically verify build and troubleshoot for errors

Acceptance Criteria:

- Given that I am a DevSecOps Engineer, when I review the CICD pipeline, then I can see scans are being conducted regularly, and then I can confirm all new code introduced into the production environment does not contain any vulnerabilities

Story Points 5

User Story 11: As a consumer, I want to confirm the data classifications are useful (real-time, accurate, and actionable) so that I can make informed decisions.

Tasks:

- Create APIs using Lambda and API Gateway to expose the data from the database to the user interface
- Design serverless Lambda functions to automatically convert PostgreSQL tables to JSON format for web UI
- Build interfaces to show both legacy and new Sector/Industry classifications in an interactive tree-map

Acceptance Criteria:

- Given that I am a consumer, when I am on the AI sector page, then I can see an interactive tree map that displays both legacy and new Sector/Industry classifications
- Given that I am a DevSecOps engineer, when I navigate to the AWS pages, then I can see that the serverless Lambda functions automatically converts PostgreSQL tables to JSON format for web UI and then I can pull new data quickly

Story Points 5

User story 12: As a consumer I want to be able to filter data by industries so that I can better target my search.

Tasks:

- Design and develop UI that will allow users to filter by industries
- Create APIs via Lambda to enable JSON output of data
- Write code to classify data by defined industries

Acceptance Criteria:

- Given I am a user, when I select a sector, then I can see all industries listed under the selected sector, and then I can drill down into any industry under the sector

Story Points 3

User story 13: As a consumer, I would like to filter data by sector so that I can better find the information I am looking for.

Tasks:

- Design and develop UI that will allow user to filter by sector
- Design and build side bar navigation
- Create APIs via Lambda to enable JSON output of data
- Write code to classify data by defined sectors

Acceptance Criteria:

- Given I am a user, when I navigate to the AT homepage, then I see all 10 sectors as individual tiles, then I select any of those sectors and I can see all selected sector information in the table below

Story Points 3

User Story 14: As a consumer, I would like to see how companies have evolved so I can better understand trends and make strategic and forward thinking decisions.

Tasks:

- Write function that will track company’s performance by industry and sector by year
- Develop UI that will show trends
- Build chart widget
- Create competitor DB table

Acceptance Criteria:

- Given that I am a consumer, when I filter for a specific company, then I can see the AT classification, Legacy classification, Market Cap, Sector/Industry weights and top competitors

Story Points 5

User Story 15: As a consumer I would like to understand the industry trends so that I can identify anomalies and potential opportunities.
Tasks:

- Pull 2018 market cap data to identify market shares

Acceptance Criteria:

- Given that I am a consumer, when I filter for a specific industry, then I can see an interactive tree map that displays both legacy and new Industry classifications, then I can see how an industry 10 years ago compares to present, and then I compare that industry to other industries in the same sector

Story Points 2

User Story 16: As the product owner I want to leverage Machine Learning / Artificial Intelligence technology so that the solution continually improves as more data is collected and result are useful to the consumer.
Tasks:

- Create corpus (parallelized)
- Create Doc2Vec corpus, a natural language processors, which learns the topics and patterns that comprise the input text
- Optimize Doc2Vec hyper parameters
- Create output vectors

Acceptance Criteria:

- Given that I am a data scientist, when I open the models.md file also posted in GitHub, then I can see all descriptions of our logic, the models, and clusters leveraged in the AT solution

Story Points 13

User Story 17: As a data scientist, I would like an interactive data science notebook solution to understand the models and data used in the solution.
Tasks:

- Setup Jupyter notebook as a serverless instance on Amazon Sagemaker
- Load in relevant python code and libraries
- Break down data analysis using descriptions and plots
- Show NLP models
- Describe each step of the AI/ML pipeline

Acceptance Criteria:

- Given that I am a data scientist, when I open the notebook, then I can see the AI/ML process from start to finish, and then I can re-run models in real-time
- Given that I am a data scientist when I open the R and Python code posted in GitHub, then I can tweak settings and parameters for analysis
- Given that I am a data scientist when I am in the AT interphase, then I can export all raw data, and then I can plot and graph of data and outputs

Story Points 8

1.1.3 Future Functionality

Alpha Taxonomy was built through an iterative process. Through our process we captured additional feathers to be included in a Product Roadmap.

As a product owner I would like additional navigation features so that the user can track their navigation path and easily drill back up.

Tasks:

- Design wireframes to include interactive bread crumbs
- Develop interphase with interactive breadcrumbs

Acceptance Criteria:

- Given that I am a user as I click into additional information, then the bread crumbs expand to track the actions I have taken, and then when I select one of the breadcrumbs, then I am directed to that section

Story points 1

As a consumer I would like to be able to identify companies that are grouped differently in the AT solution than their self-selected SEC 10-k filings.

Tasks:

- Develop R Python code to compare and track discrepancies
- Design wireframes to include a notification of discrepancies
- Develop UI to include notification of discrepancies

Acceptance Criteria:

- Given that I am a consumer, when I select an industry, then all companies listed underneath, then I see a flag on companies that have discrepancies in model and sec 10-k filings

Story points 3

As a consumer I would like to be able to validate classification via the interphase.

Tasks:

- Develop R Python code to pull natural language that describes companies classification
- Design wireframes to include a company pop-up with justification for company’s classification
- Develop code to build a company pop-up that includes justification for company’s classification on the UI

Acceptance Criteria:

- Given that I am a consumer, when I select the company name, then a pop-up appears with a justification for why that company belongs to that industry

Story points 8

guidehouse.com