Solution.PDF Submission

Guidehouse, LLP, Data Universal Numbering System (DUNS): 079529872, Solicitation Number #: 70SBUR19Q00000066, Shannon White, Partner, 1800 Tysons Boulevard, 7th Floor, McLean, VA 22102-4257, Telephone: (571) 296-2571, Email: swhite@guidehouse.com

The objective of the Solution.PDF document is to provide the government with additional testimonies and artifacts that illustrate our design approach, data configuration, Machine Learning and cluster naming methodology.

Our Solution

The Alpha Taxonomy solution provides the ability to ingest massive quantities of publicly traded company data, normalize the data into an interpretable format by cleansing and standardizing, storing the data with clustering capabilities, and executing Artificial Intelligence (AI) and Machine Learning (ML) algorithms for data insights.

Create and Connect

Day 1 the Guidehouse team convened to conduct an ideation and discovery session using a Human Centered Design approach. We began by understanding the problem we were trying to solve for: **Publicly traded companies are not effectively categorized into sectors or industries**. **In order to remain competitive companies are forced to adapt, grow, and even alter their services offerings to stay relevant. As a result, sectors and industries are constantly transforming, making it difficult for analysts, regulators, financial advisors, and companies to classify companies by sectors or industries, compare companies, identify trends, understand the current landscape, and anticipate change.**

As part of our discovery and ideation efforts Guidehouse identified our core users, developing personas. We started designing the personas by asking the question: "who are we designing for?" From there we identified our two core users: the data scientist and the consumer – the consumer represents several professionals including economist, company executive, investors, and US publicly traded company employee. Next, we empathized with the users to uncover the common behaviors, hidden motivations and influences, pain points, and bright spots of the users. We then logged the motivations and frustrations of the users into our personas, as shown in the image below, and leveraged the personas to drive the solution and user story development.

Veronica

#datascientist

JOB TITLE
Data Scientist

AGE
21+

HIGHESTLEVEL OF EDUCATION
Doctoral or Professional

GOALS & MOTIVATIONS
- Wants to understand the output data at a deeper level
- See the raw data and weightings
- Needs help making models
- Continuously improve and refine my models

FRUSTRATIONS
- Lack of visibility into the classification of raw data
- Confusing models without documentation
- Current classifications are hard to interpret and inflexible
- Automation isnot leveraged

Rob, Chris, Will & Sam

#consumers

JOB TITLES
Economists
Company Executives
Investors
Employee

AGE
21+

HIGHESTLEVEL OF EDUCATION
Doctoral or Professional

GOALS & MOTIVATIONS
- Pinpoint industry investment opportunities across developed and emerging markets
- Get information on a certain company's competitors
- Capture and assess the impact of global, regional or local industry trends on a portfolio
- Penetrate new markets
- Wants to trust the data models

FRUSTRATIONS
- Poor classifications of companies, sectors, and industries
- Outdated or subjective classification schemes
- Human error due to manual classification processes
- Lack of access to solid data (due to subscription-based products, etc.)
- Sometimes struggle to trust the data they have access to

We created a vision for what our product would accomplish: **accurately classify companies, traded on the three major U.S. financial securities exchanges and corporations, into 10 sectors and 100 industries**. Team Guidehouse's Alpha Taxonomy (AT) is a cloud based AI/ML analysis solution leveraging two primary unsupervised machine learning techniques, one for Natural Language Processing and one for Hierarchical Clustering, to define and classify sectors and industries based on current and historical data. Our solution provides the user with a way to effectively access the results to help the user to make informed and strategic decisions.

Iterate

Our team took an iterative approach to our design and adopted a cyclic process of prototyping, testing, analyzing, and refining our product to best meet the end-users' need. Leveraging the personas and user stories our team built wireframes to help visualize the user's journey. We developed the wireframes on the InVison tool, which can be viewed at https://guidehouse.invisionapp.com/share/JNS40U7EFC2#/screens.

Throughout each sprint, the team conducted usability tests—A/B, Think Out Loud, Man-on-the-Street and Remote—to challenge our assumptions, iterate on the design and content, and validate the original value proposition. Feedback from these testing sessions was shared among the team during scrum. The team adjusted planning and priorities based on the testing results.

As the design evolved, so did the codebase. The development team rapidly added new code to support functionality continuing to iterate on and mature existing code. During each sprint the code was subject to three levels of testing rigor—unit tests, expert reviews and peer reviews—with defects corrected at the time of discovery.

We built the solution using the infinity HTML scroll to provide users access to large amounts of information with minimal clicks and navigation steps.

Alpha Taxonomy was created and built with the techniques and corresponding benefits documented below:

General Architecture and Design Criteria	Technical Approach and Innovative Techniques	Derived Benefits
Fully Automated Open Source Data Ingestion Pipeline	A fully automated system that leverages APIs and web scrapers to continuously pull both current and historical company data from a variety of sources (SEC 10K filings,	Completely automated and repeatable data collection process Runs multiple requests in parallel cutting data

	corporate web sites, Wikipedia, Yahoo Finance). The data collection stack is built on the R programming language and implemented on our DevSecOps microservice architecture.	collection time from days to minutes. Automatically infers and pulls the most relevant information from websites, Wikipedia, and documents with no human intervention
Use advanced algorithms: Natural Language Processing	Leveraged purely publicly available open source data to drive the model – reading from massive bodies of documents, web, and text to derive insights at scale. Used the state-of-the-art natural language processing method Doc2Vec to embed documents into a numeric vector space such that similar documents are closer to each other	Mapping to numeric vector space allows user to use various machine learning techniques reserved for numeric data only. Completely automated and repeatable NLP process to pull new insights as open source data is updated. Enabling a computer to actually understand context from dissimilar heterogeneous data sources.
Use advanced algorithms: Artificial Intelligence/Machine Learning	Used unsupervised hierarchical clustering to create groups of companies based on numeric representation of the documents. Creates an intuitive picture of company relationships to show actual hierarchies between sectors and industries. Developed custom pattern matching techniques to actually have a computer come up with names for clusters.	Business problem's requirement of all industries bubbling up to sectors is sufficed due to the tree-like structure. Our unique purely unsupervised approach enabled finding new classifications without the onerous task of labeling thousands of data elements by hand.
Read/write data via Serverless Application	Each service exposes its capabilities via a set of clean	Eliminates the need for costly backend servers by process

Programming Interface (API)	API's implemented entirely through a **serverless** Lambda and API Gateway backend. AWS API Gateway service proxies all requests to backend services. The cost incurred by a serverless application is based on the number of function executions, measured in milliseconds instead of hours.	agility -- Smaller deployable units result in faster delivery of features to the market, increasing the ability to adapt to change. Cost of hiring backend infrastructure engineers goes down and reduced operational costs. Reduced liability, no backend infrastructure to be responsible for and zero system administration. Easier operational management. Fosters adoption of Nanoservices, Microservices, SOA Principles. Faster set up. Scalable, no need to worry about the number of concurrent requests. Monitoring out of the box.
Fully automated ML based ETL process	After successful completion of the AI/ML process the data is sent to an AWS S3 storage bucket. This then automatically triggers scripts within AWS Glue that spin up. Spark instances to transform and load the data into our AWS RDS PostgreSQL database.	Relationships and mappings within the data are automatically inferred and cleanly ingested into the appropriate sector, industry, and company data models without any human intervention.

US Government Web Design Standards (https://standards.usa.gov/)	Implemented the U.S. Web Design System for base styling and some functionality.	Simplicity and consistency across Government services.
Built on the principles of TDD and BDD	Followed test first approach from the user story through the development of code. Create Agile format acceptance criteria to derived common team understanding. Agile code and tests are contained in our code repository.	Quality code and 100% test coverage. Easily verified user stories with less misunderstanding of requirements. Zero vulnerabilities. Accurate estimation and targeted testing speeds overall team velocity.
Satisfy the business needs and offer a good user experience	Followed UI design principles from the US Government standards and used the templates for UI design markups. Built the quality into the pipeline for a zero defect, zero down time deployment.	Easy look and feel web applications with a targeted 99.99% up time.
Meet DHS Section 508 requirements	Automated accessibility testing for 508 requirements in the pipeline using Axe-core/Deque. Highly configurable, automatically determines which rules to run based on evaluation context.	508 compliant application, zero false positives and better use of 508 testers for exploratory testing. Addresses 508 compliance in future releases with tests as acceptance criteria as part of user stories.
Automated Unit test scripts and Functional testing including exception handling	The CI pipeline implements automated Unit tests, functional testing, security and quality scans along with building and	Increased effectiveness and efficiency and code coverage. Built in quality.

	scanning containers for security vulnerabilities.	
Containerized and microservices-based architecture and deployed on the OpenShift Enterprise Container Platform	Our approach leverages AWS Lambda serverless compute as well as OpenShift containers to encompass all of the patterns of a modern microservice architecture including service discovery, circuit breaker, intelligent routing, and client-side load balancing.	Proven tool stack and framework that delivers scalability and reliability. Rapid delivery that allows developers to hit the ground running and become productive in short order.
Demonstrate best practices for microservices architecture deployed on OpenShift	Terraform OCP tooling is leveraged for automated deployment and management of services. Terraform OCP secrets for protection of sensitive data. Terraform OCP templates for recreating entire configuration are version-controlled. Dynamic metrics-based scaling Microservices employ Bounded Context with data separation.	Zero down time deployments. Templates ensure recreation of Terraform OCP objects on-demand. Highly secure and resilient system, which optimizes resource consumption. Dynamic scaling based on real-time metrics ensures the right capacity at the right time.
Present a fully-automated CI/CD pipeline with appropriate gate controls	Implements automated CI/CD pipelines using pipeline as code. GitHub Flow implementation enforces quality gates even before code is merged, the pipelines implement completely automated testing in the pipeline.	Agile DevSecOps team makes code changes with confidence. Secure and quality code promotions.

Data collection

Our solution provides a robust data set to the end user. During the conceptualization we conducted an environmental scan to identify what information was available and useful to our user. We leveraged R and Python to collect data from several sources, including raw data from the SEC EDGAR 10-K, company websites, Wikipedia pages and Yahoo Finance. Our data ingestion model achieves a superior coverage ratio of 98% of companies traded on the three major U.S. exchanges.

Our Extract, Transform & Load (ETL) architecture is fully automated and deployable with two commands. The solution can be configured to retrieve more up to date source data via a scheduler, using an AWS Lambda conditional trigger, or manually by running a single command in the AWS Console. The ETL architecture is also parallelized with AWS Lambda so the data can be retrieved within hours.

Artificial Intelligence /Machine Learning (AI/ML)

What sets our solution apart was our unsupervised machine learning technique. We leveraged Doc2Vec which is a state-of the art application, published by Google in 2014, to convert the data collected into numbers, which we then applied hierarchical clustering that groups or separates objects based on shared or dissimilar characteristics to help define industry and sector clusters. In addition, we leverage the additional AI algorithms, the Levenshtein approach and the Bag of Words (BoW) approach and to name and classify clusters using natural language.

Team Guidehouse took an iterative approach to naming the clusters, initially our team scanned the 10-K documents and other web-scrapped descriptions about the companies to identify the most "important" words for each cluster. However we determined this approach needed to be further refined, as the results were too broad and many of the "important words" did not derive significant meaning. After running several experiments, we found that using a bag of words (BoW) approach worked the best. For each sector cluster we put all the words from the SIC descriptions of all the companies in a bag – meaning we discarded all information about the order or structure of words. We then selected the top three words with the highest numbers of occurrences as the constituents to name the cluster. To make the sector names sound more natural and intuitive, we introduced a small set of grammar rules to guide the ordering of the three words based on nouns or adjectives.

Validating our approach:

Although our clustering algorithm may cluster companies in a manner different than the existing approaches, we consider companies that self-classify themselves by using the SEC SIC codes a reliable measure for validating clustering results, as companies with the same SIC code should see a good concentration in clusters. We generated multiple clustering sector/industry groups by using different text descriptions as input, including the 10-K documents of all publicly traded companies that can be downloaded from the SEC's EDGAR web site between 2008 and 2018, the 10-K documents of all the companies for 2018, the 10-K documents for 2018 along with text

on companies' websites and in their Wikipedia pages, and a few other combinations. It turned out using text descriptions from many sources did not result in better clustering results than using a single source of text input. Therefore, we used only 10-K documents of all the companies for 2018 as the base for generating sector/industry clusters. To validate clustering results, we wrote scripts to generate a ranked list of SIC codes for each sector/industry cluster, and calculated the percentages for the different SIC codes in a cluster and the accumulated percentages for the top SIC codes in the ranked list. The results are automatically generated, please see results in table below. By so doing, we were able to examine whether the companies with the same SIC codes might form groups in clusters.

Sector Names
Pharmaceutical and Biological Preparations
Surgical Instruments and Apparatus
Services-Prepackaged Services and Software
Other Stores and Products
Devices Related Semiconductors
Equipment Machinery And Products
Banks Commercial And State
Real Insurance And Investment
Natural Gas And Petroleum
Real Estate And Investment

Industry Names
Pharmaceutical Preparations and Electromedical & Electrotherapeutic Apparatus
Biological Products (No Diagnostic Substances) and Medicinal Chemicals & Botanical Products
Biological Products (No Diagnostic Substances) and Services-Commercial Physical & Biological Research
Biological Products (No Diagnostic Substances) and **Services-Commercial Physical & Biological Research**
Arrangement Of Transportation Of Freight & Cargo a**nd Hazardous Waste Management**