clearscale

aws partner network

Premier
Consulting Partner

Implementing a Large AWS Data Lake for Analysis of Heterogeneous Data

C4ADS
innovation for peace

Executive Summary

C4ADS (Center for Advanced Defense Studies) is a nonprofit organization based in Washington DC that is dedicated to providing data-driven analysis and evidence-based reporting on global conflict and transnational security issues. In this pursuit, C4ADS focuses on a variety of issues, including threat finance, transnational organized crime, and proliferation networks.

> “Our partnership with AWS has been a great experience, and the platform has proven to be a powerful tool in storing and piecing together disparate data sets across different languages and format. ClearScale provided us the boost we needed to address issues of transnational security in a more timely fashion.”
>
> **David W. Lynch,** Chief of Analysis

The Challenge

The world is a complex ecosystem of people, economies, competing interests, and political ambiguity. Being able to track many different events to determine if there are patterns that would warrant a more critical look and analysis is a difficult task, even under the best conditions. With new regional or political developments each day, sometimes even hour by hour, combing through enormous sets of data is challenging; especially when that data is from different sources and in various formats.

C4ADS is tasked with just this sort of activity. Their clients require evidence-based and data-driven analysis concerning global conflict and transnational security issues. With a focus on identifying the drivers and enablers of such conflict, this organization has to be absolutely confident in the analysis and assessments they provide. However, the first step to performing any sort of review requires analysts to comb through extensive records from different sources and formats to compile a list of potential hits.

As C4ADS increased the number of datasets it ingested, new challenges arose, specifically the ability to make use of all the data at its disposal. As more and more data has become available, their analysts were finding it difficult to sift through all of the incoming information in a quick and expedient way. The company approached ClearScale, an AWS Premier Consulting Partner, and wanted to see if there was a way that they could leverage what they did currently by using AWS to assist in making the data more user-friendly.

The ClearScale Solution

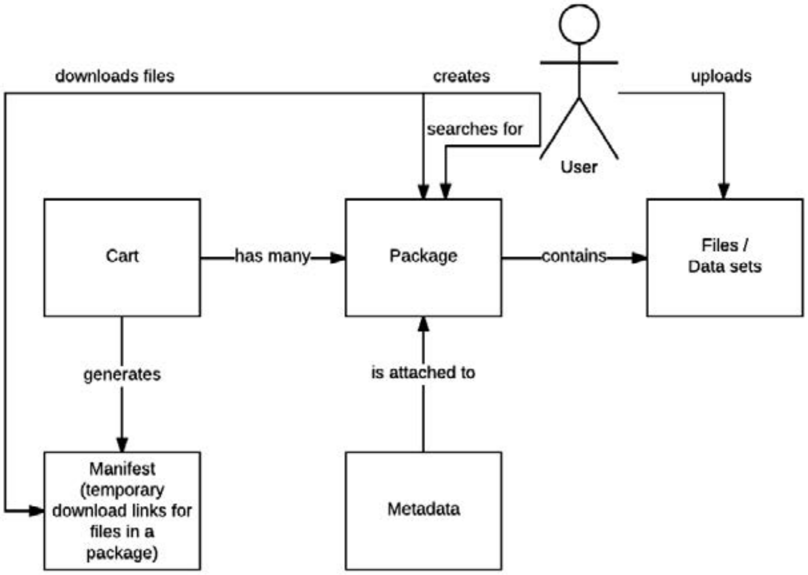
The challenge put forth by C4ADS was that a solution had to be implemented quickly, provide the ability to scale as needed, and be extremely secure given the nature of the information they were reviewing. With these three criteria in mind, ClearScale reviewed various designs and approaches that they could develop and implement on AWS.

Data Storage with Data Lake Approach

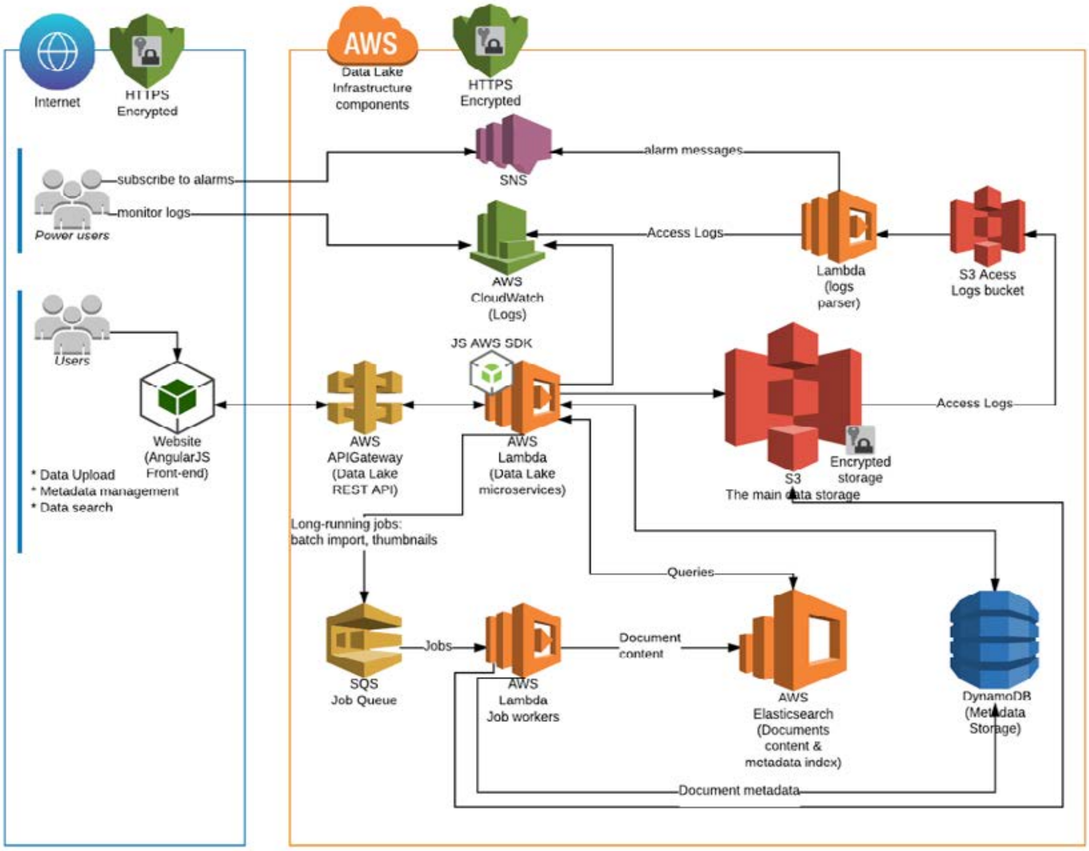
The biggest challenge was finding a way to aggregate multiple different file formats (such as PDFs, emails, Microsoft Word and Excel files, logs, XML and JSON files) while still allowing C4ADS to perform easy searches within a large data repository. It rapidly became clear that to accomplish the requirements laid out by the client, ClearScale would have to implement a Data Lake approach within an AWS Virtual Private Cloud (VPC). Unlike traditional data warehouse methodologies that require data to conform to a specific set of schema, a data lake allows for any number of data types to be stored and referenced, so long as those data types have a consistent approach to querying and retrieving data.

It was immediately clear that trying to collapse or conform all the various file types that were available into a normalized format would be too resource-intensive. To overcome this, ClearScale chose instead to implement a solution that would tag all uploaded file content with consistent metadata tagging which, in turn, would allow for greater visibility and speedier search results. This automated metadata tagging for each file that was uploaded either manually or via bulk upload would mimic the client's existing folder structure and schema that they had adopted internally. This approach would ensure that the new solution would be easily understood by analysts that were already familiar with the current operational processes.

Data Flow Model



System Architecture Diagram



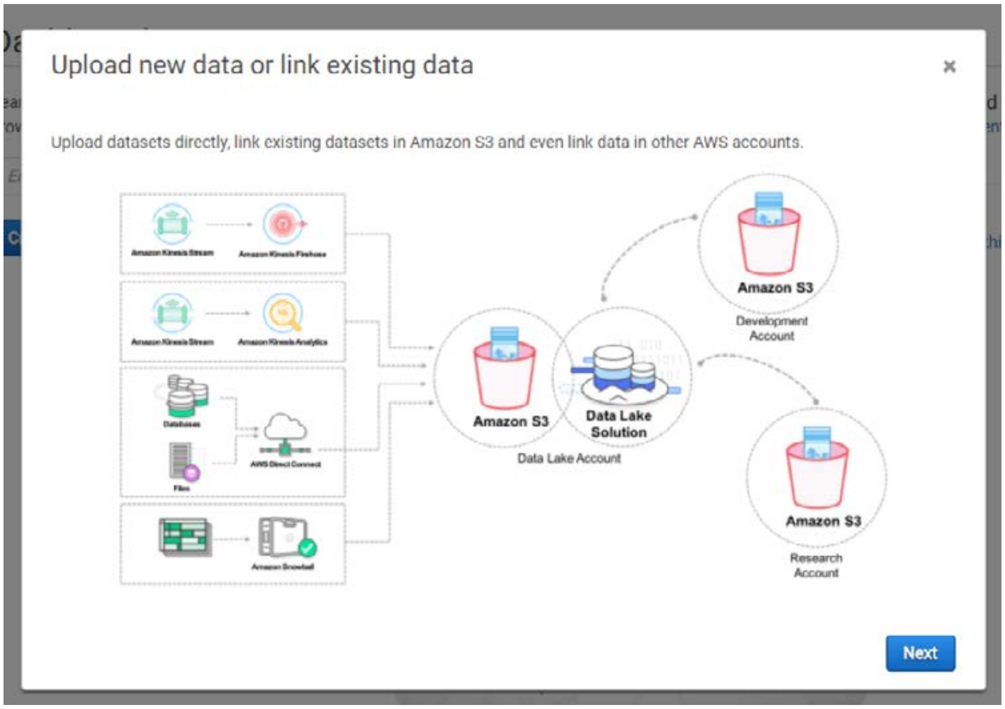
Web-Based User Interface (Web UI)

To access and search these records, ClearScale designed and implemented a web-based user interface. This UI was designed to allow for complete management of the data sources — including data upload — beyond simply searching the Data Lake. From a data repository perspective, ClearScale needed to build and deploy a solution that was scalable and reactive to increased demand but also highly secure. To accomplish this, a combination of AWS S3 was used for the storage of the data uploaded, and DynamoDB for the storage of the file metadata; ElasticSearch was used for the robust search querying that was required.

In order to get the data uploaded, ClearScale leveraged AWS Lambda and API Gateway services to properly ingest the data and automate the creation of the file metadata. Both CloudWatch and CloudTrail were also put in place to monitor resource usage and serve as triggering mechanisms to scale the environment as required.

The entire solution was encased in AWS VPC for robust security and Cognito for SAML based authentication. This approach guarantees that the information was behind a robust security layer with additional work done for data to be encrypted both at rest and in transit. It also insured that administrators could grant access to specific document types based on group roles, both for internal and external role types.

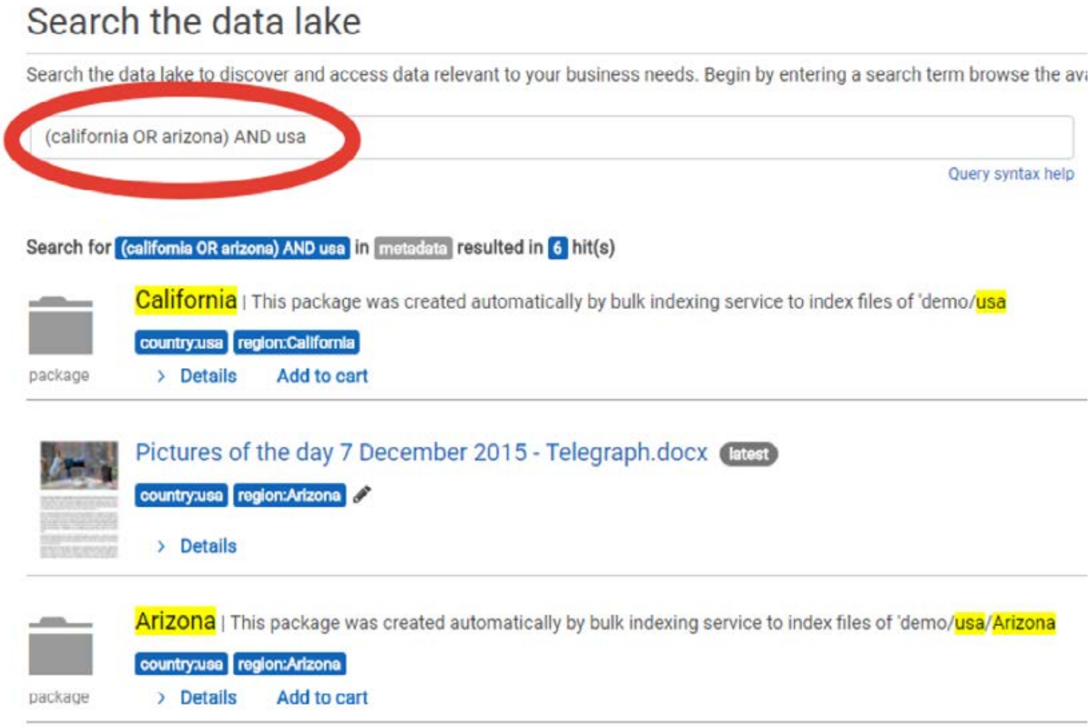
UI Welcome Screen



Upload new data or link existing data

Upload datasets directly, link existing datasets in Amazon S3 and even link data in other AWS accounts.

UI Search



Search the data lake

Search the data lake to discover and access data relevant to your business needs. Begin by entering a search term browse the ava

(california OR arizona) AND usa

Search for (california OR arizona) AND usa in metadata resulted in 6 hit(s)

California | This package was created automatically by bulk indexing service to index files of 'demo/usa

country:usa region:California

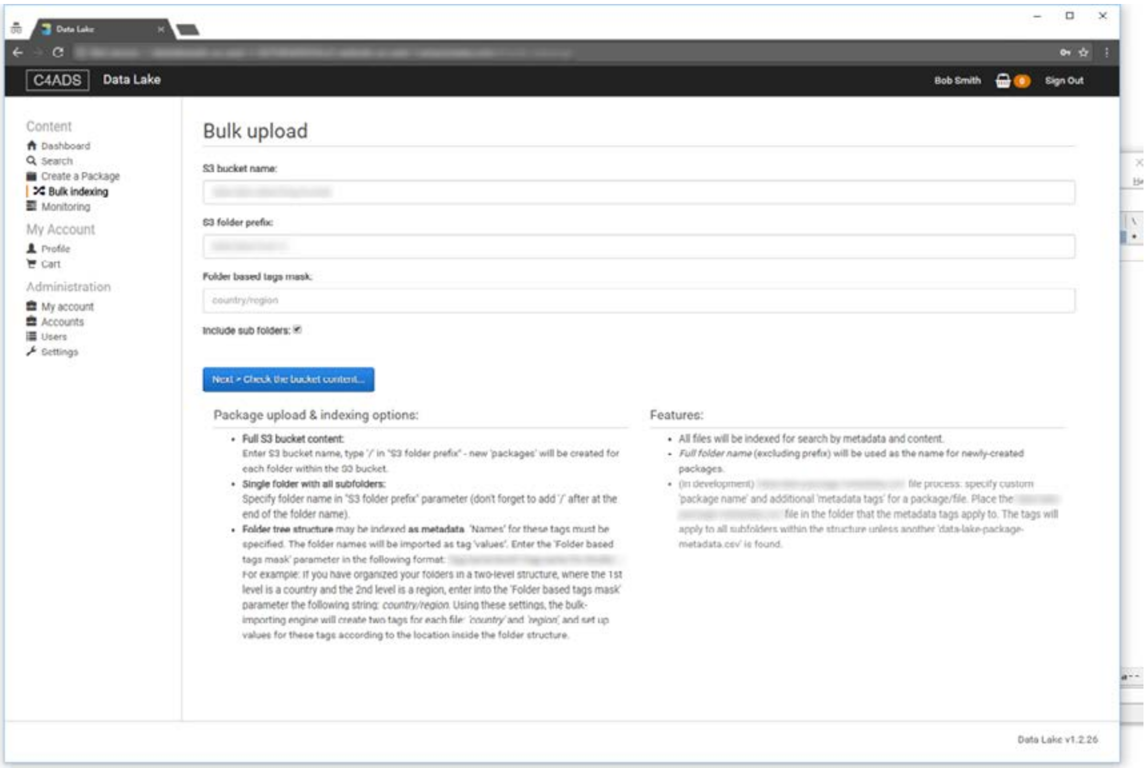
Pictures of the day 7 December 2015 - Telegraph.docx latest

country:usa region:Arizona

Arizona | This package was created automatically by bulk indexing service to index files of 'demo/usa/Arizona

country:usa region:Arizona

Bulk Indexing — Add and Index an existed S3 Bucket or Folder



Bulk upload

S3 bucket name: [redacted]

S3 folder prefix: [redacted]

Folder based tags mask: country/region

Include sub folders: ☑

Package upload & indexing options:

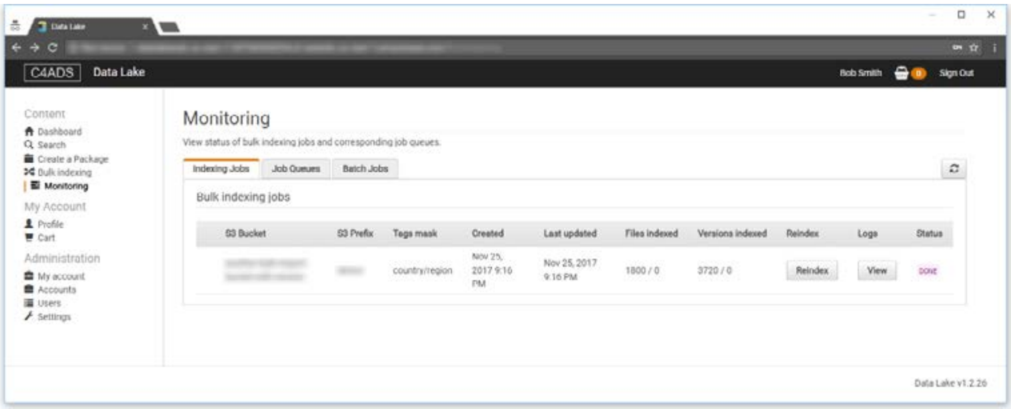
- **Full S3 bucket content:**
  Enter S3 bucket name, type '/' in "S3 folder prefix" - new 'packages' will be created for each folder within the S3 bucket.
- **Single folder with all subfolders:**
  Specify folder name in "S3 folder prefix" parameter (don't forget to add '/' after at the end of the folder name).
- **Folder tree structure** may be indexed **as metadata**. 'Names' for these tags must be specified. The folder names will be imported as tag 'values'. Enter the 'Folder based tags mask' parameter in the following format: [redacted]
  For example: If you have organized your folders in a two-level structure, where the 1st level is a country and the 2nd level is a region, enter into the 'Folder based tags mask' parameter the following string: *country/region*. Using these settings, the bulk-importing engine will create two tags for each file: *'country'* and *'region'*, and set up values for these tags according to the location inside the folder structure.

Features:

- All files will be indexed for search by metadata and content.
- *Full folder name* (excluding prefix) will be used as the name for *newly-created* packages.
- (in development) [redacted] file process: specify custom 'package name' and additional 'metadata tags' for a package/file. Place the [redacted] file in the folder that the metadata tags apply to. The tags will apply to all subfolders within the structure unless another 'data-lake-package-metadata.csv' is found.

Data Lake v1.2.26

Bulk Indexing — Monitoring of Long Time Backend Tasks



Monitoring

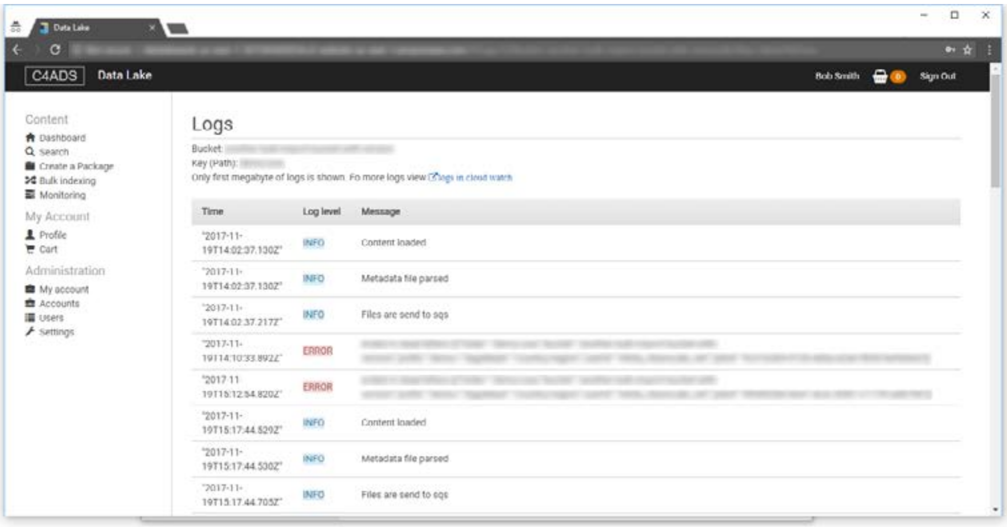
View status of bulk indexing jobs and corresponding job queues.

Bulk indexing jobs

S3 Bucket	S3 Prefix	Tags mask	Created	Last updated	Files indexed	Versions indexed	Reindex	Logs	Status
[redacted]	[redacted]	country/region	Nov 25, 2017 9:16 PM	Nov 25, 2017 9:16 PM	1800 / 0	3720 / 0	Reindex	View	DONE

Data Lake v1.2.26

Bulk Indexing — Login and Automatic Errors Handling

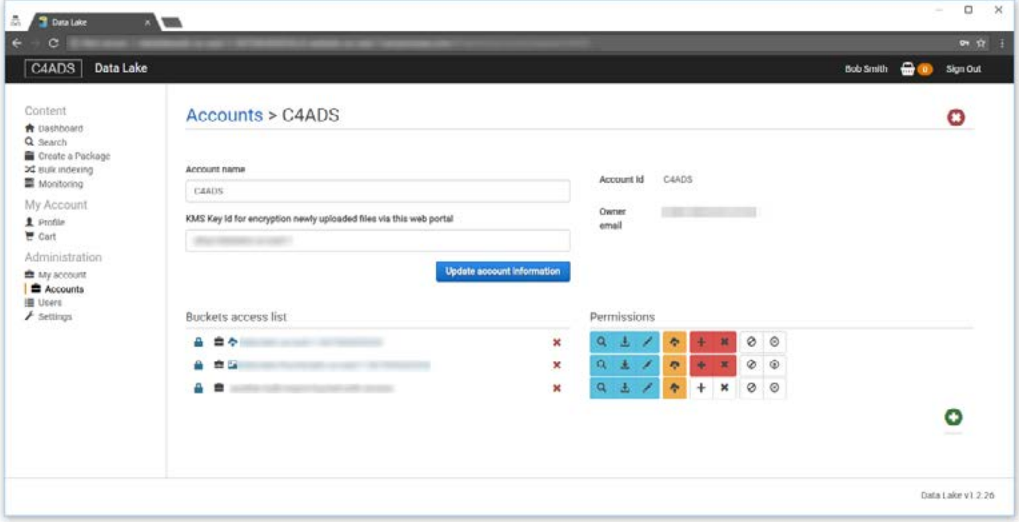


Logs

Bucket: [redacted]
Key (Path): [redacted]
Only first megabyte of logs is shown. Fo more logs view logs in cloud watch

Time	Log level	Message
"2017-11-19T14:02:37.130Z"	INFO	Content loaded
"2017-11-19T14:02:37.130Z"	INFO	Metadata file parsed
"2017-11-19T14:02:37.217Z"	INFO	Files are send to sqs
"2017-11-19T14:10:33.892Z"	ERROR	[redacted]
"2017 11 19T15:12:54.820Z"	ERROR	[redacted]
"2017-11-19T15:17:44.529Z"	INFO	Content loaded
"2017-11-19T15:17:44.530Z"	INFO	Metadata file parsed
"2017-11-19T15:17:44.705Z"	INFO	Files are send to sqs

Multi-tenancy — Agile Access Setup



Accounts > C4ADS

Account name: C4ADS

KMS Key Id for encryption newly uploaded files via this web portal: [redacted]

Account Id: C4ADS

Owner email: [redacted]

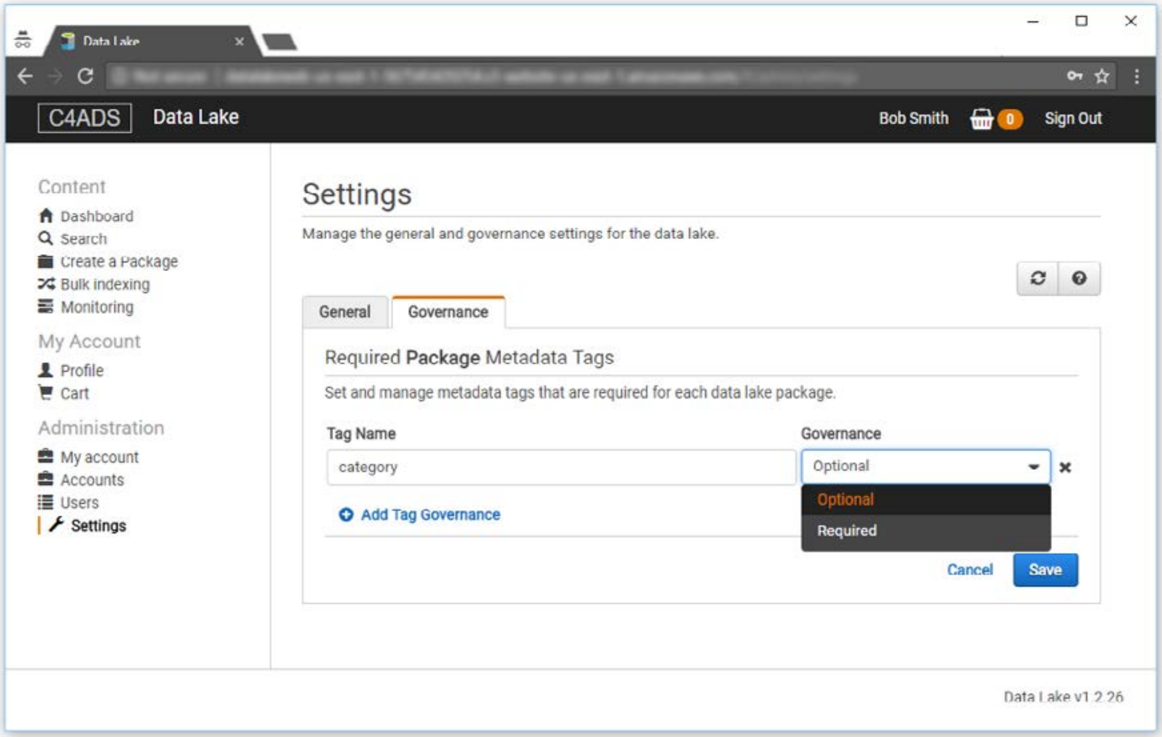
Buckets access list

- [redacted]
- [redacted]
- [redacted]

Permissions

Data Lake v1.2.26

Metadata — Governance



Settings

Manage the general and governance settings for the data lake.

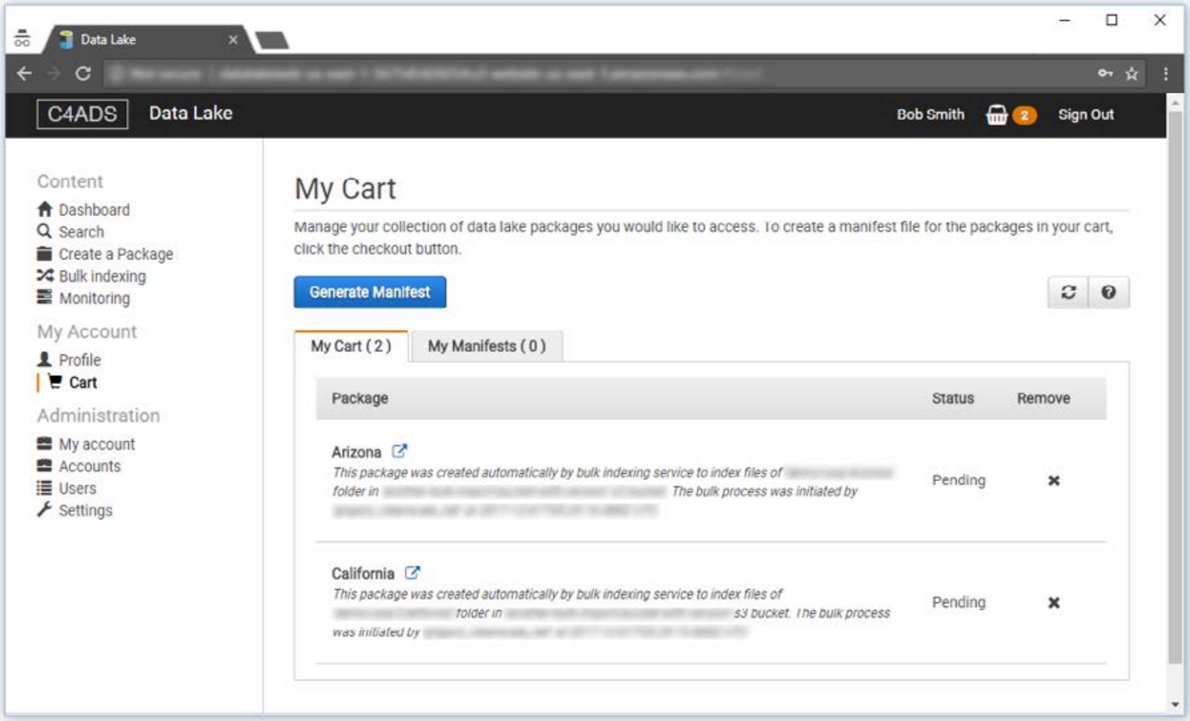
Required **Package** Metadata Tags

Set and manage metadata tags that are required for each data lake package.

Tag Name	Governance
category	Optional

Data Lake v1.2.26

Cart — Storing and Exploring Results in Personal Cart



My Cart

Manage your collection of data lake packages you would like to access. To create a manifest file for the packages in your cart, click the checkout button.

Package	Status	Remove
Arizona *This package was created automatically by bulk indexing service to index files of [redacted] folder in [redacted] The bulk process was initiated by [redacted]*	Pending	✖
California *This package was created automatically by bulk indexing service to index files of [redacted] folder in [redacted] s3 bucket. The bulk process was initiated by [redacted]*	Pending	✖

The Benefits

The turnaround time from design to delivery to C4ADS was a mere two months, including deployment of the solution in both a Staging and Production environment as well as training for C4ADS staff on how to use the new solution. The first release provided everything that C4ADS originally asked for: it had to be deployed quickly, it had to have the ability to scale as needed, and it had to be highly secure. Launched in October 2017, the solution has already optimized the analysts' job activities by giving them the tools necessary to do wide-ranging search profiles and aggregate disparate heterogeneous data types.

Next Phase

Later releases will introduce more robust security measures that will allow C4ADS to extend the service out to their partner organizations. It will also provide multi-lingual support and optical character recognition (OCR) technology to aid in identification of important data markers in the data that is uploaded.

There are plenty of challenges in the business and technology landscape. Finding ways to overcome these challenges is what ClearScale does best. By bringing our own development resources to bear on these complex problems, we can design, build, test, and implement a solution in partnership with your organization, thus allowing you to focus on more pressing matters in running your day-to-day operations.