collibra®

Whitepaper

Elevating Your Cloud Strategy with AWS and Collibra

collibra.com

As an organization begins their digital transformation, the cloud often becomes a focal point of that migration. Moving from an on-premise infrastructure to a cloud or hybrid cloud environment can be a daunting task with a lot of room for error; couple that with upwards of hundreds of thousands of sensitive personal data records, and that error becomes a very real danger to an enterprise's operation. It's critical to make sure that all of your information, including your data lake, is migrated properly and efficiently, which makes a sound cloud migration strategy key. This is where Collibra's partnership with Amazon Web Services (AWS) can help.

Though brief, the history of data lakes has gone through a few remarkable changes. The term "data lake" was once synonymous with a centralized on-premise Hadoop cluster, however, the speed of data acquisition has forced organizations to move to a distributed setup. And when the data is acquired faster than it can be moved, a centralized data lake is of no avail.

The solution to this issue: the cloud. Cloud services, specifically AWS, offer the flexibility and scalability organizations need to store their data lakes without using the amount of resources on-premise storage requires. The key is balance. As organizations continue to search for the right balance of their cloud and on-premise solutions, they need to consider the challenge of separating the movement and location of their data from the need and experience of how their users find it.

Finding Confidence When Moving to the Cloud

Getting a single view of all of an organization's data and analytics assets is a challenge. There are several data dimensions that can clutter the overview, making it difficult to find the most accurate, trustworthy data. For this reason, a catalog is a critical asset because it gives you a comprehensive view of your distributed data landscape. Data in your catalog should be served to you in the same way, independent of its source system, location, or ownership silos.

An overview of this scope has its own challenges to consider: How can you trust data that is coming from sources you're not familiar with? A catalog that simply enables you to find data isn't sufficient; you need a catalog that indicates whether data is trustworthy or questionable. A proficient catalog will break your data out of its silos and help you determine which data you can use and for how long in an easy-to-understand fashion.

Additionally, you need a catalog that can provide the business context of your data. Business context enables your business users to have more insight into your data, in turn, making them more inclined to use the data which will generate business value.

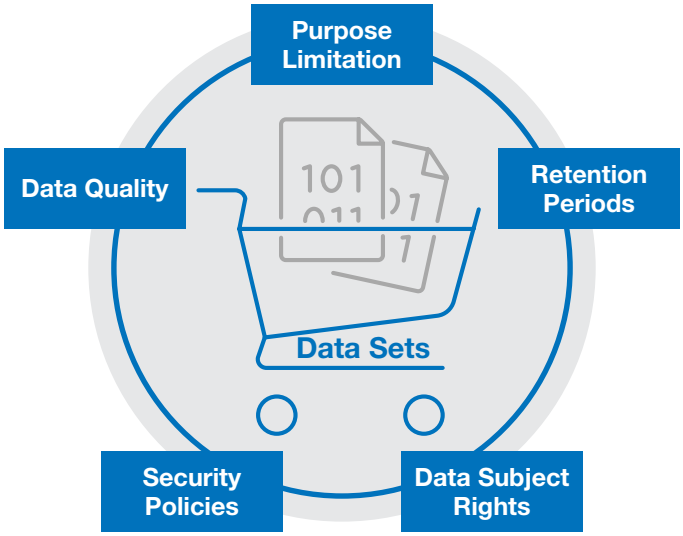
AWS and Collibra: A Solution to the Dilemma

Virtually every data platform has a version of a technical data catalog. In Amazon's case, their Glue service crawls the data lake for technical metadata. While AWS Glue is a premier service for technical users, it doesn't give business users the information they need like business context, personnel context, objective quality metrics, subjective appreciation, lineage, and internal and external data usage policies.

Collibra has been working with AWS since Collibra was founded. All of the cloud instances Collibra sells today are hosted on AWS and our collaboration with AWS is continuing to an official partnership. This partnership will include moving specific core components of the Collibra platform to AWS as well as leveraging AWS Glue and other services. For customers who have a data lake built on the Amazon stack, Collibra added support for the AWS Simple Storage Service (S3) in addition to the support for Amazon Redshift that's already in place. This means that customers who currently have their data lake on AWS S3 can now use Collibra Catalog to understand the business context of their data in a governed state.

Amazon is known for its online shopping service and now the Collibra Catalog offers a similar shopping capability for data. Collibra's partnership with AWS provides our customers with a one-stop shop for data. The Collibra Catalog allows you to track the entire data shopping process including:

- Guidelines and information regarding the data usage
- Awareness of the retention periods and ability to revoke access when the period has expired
- Indication of data quality
- Enforced security policies
- Data subject tracking



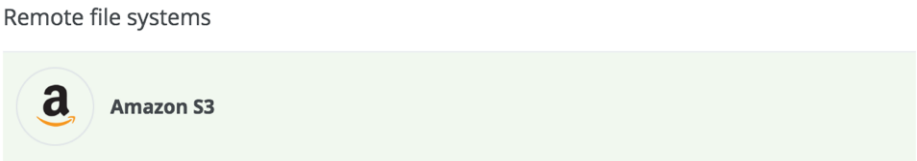
The Collibra Catalog gives you a solution to visualize the business context of your entire data landscape by providing connections to a set of BI tools and to over 300 data and analytics tools and applications.

Cataloging Amazon S3 Content with Collibra

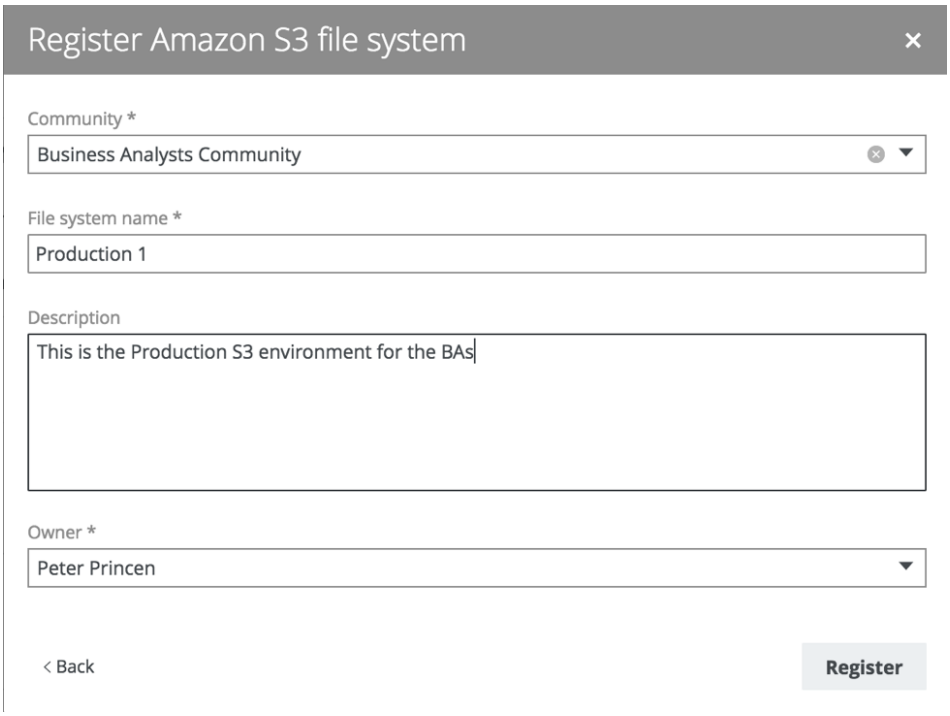
Our integration with Amazon S3 uses the AWS Glue service to complete tasks like crawling and profiling your data. Traditionally, these tasks require a lot of resources to complete; however, using AWS Glue, those resources are only deployed at the time they are needed which generates enterprise-wide efficiencies.

How it Works

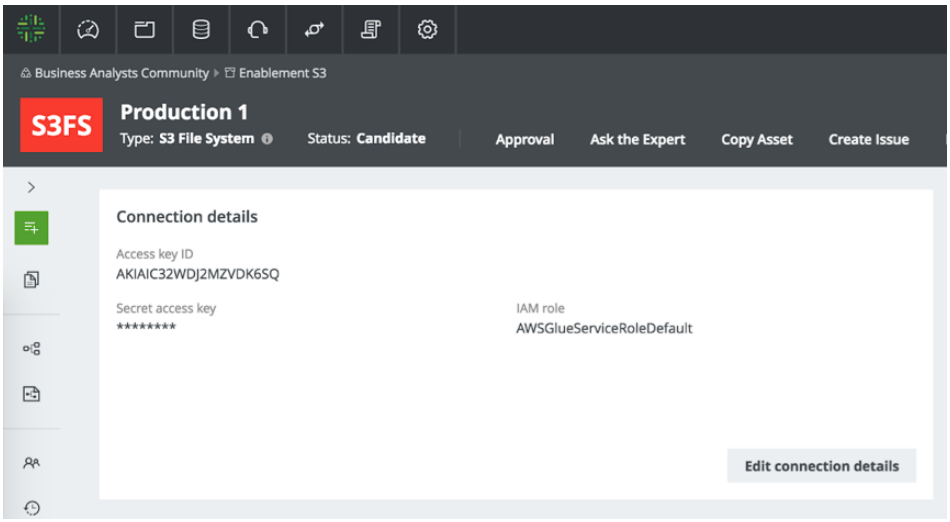
In Collibra Catalog, indicate that you would like to register a data source, then click on "Amazon S3."



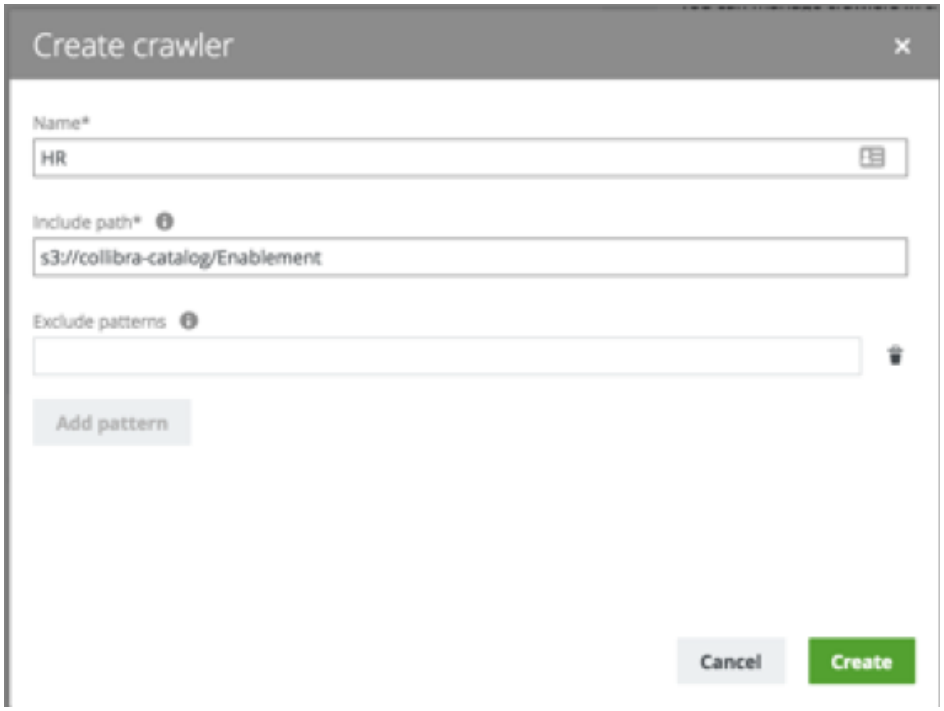
Give the file system a name and determine where you want to store the results of the crawling and who should own this file system asset in Catalog. You can also provide a description for the file system, though this isn't required.



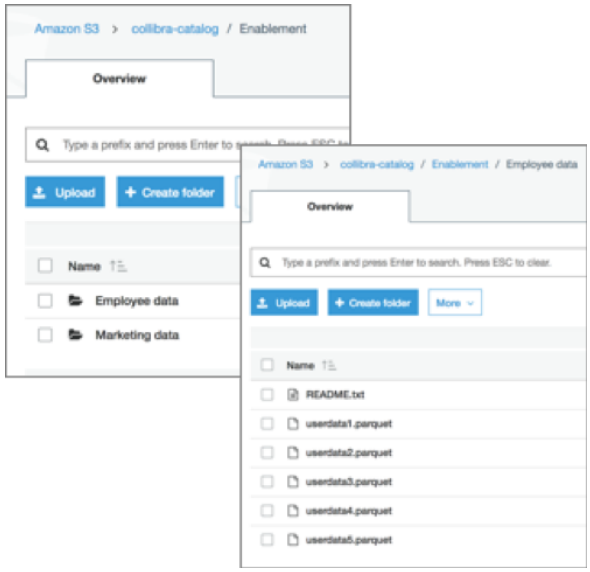
After clicking "Register," the file system asset is created and you can provide the AWS credentials required to crawl the file system. Then, enter the access key, the secret key, and the name of an IAM role to allow the use of the Glue service and gain access to the buckets you would like to crawl.



Once you are connected to the file system, you can begin to define the areas of the file system you want to register in the catalog. To do this, you need to define one or more crawlers. To define a crawler, you need to give it a name and determine which folders to include or not include. This can include folders, certain file types, files with a particular naming convention, etc.

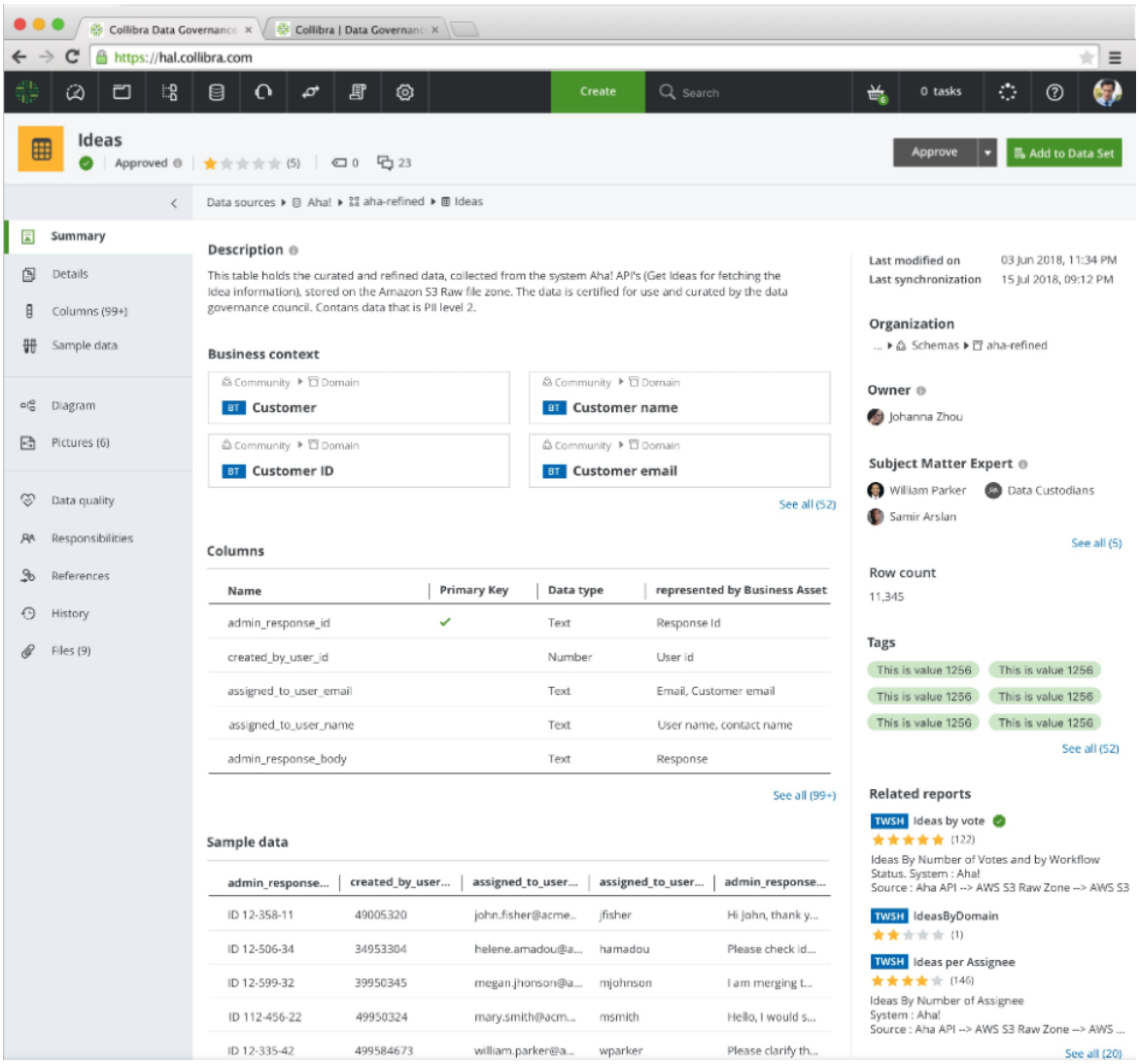


In the example below, the Enablement folder contains two subfolders. Both will be registered in Collibra Catalog. In this situation, the crawler recognized the different files in the Employee Data folder as a single parquet file that was split for efficiency purposes; thus, it is represented as a single table within Catalog.

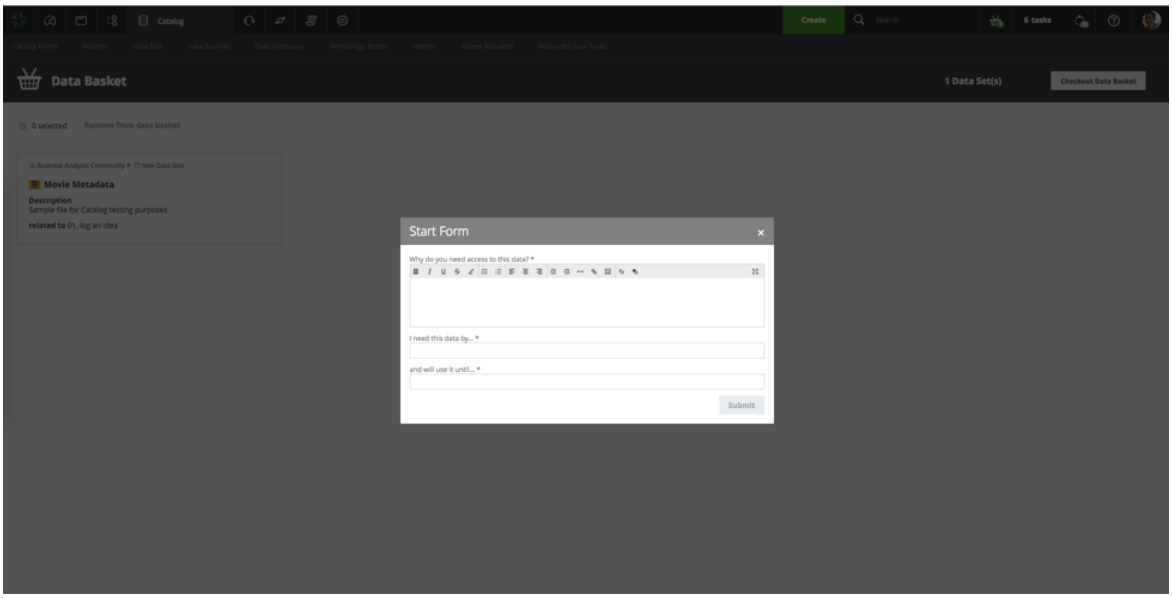


Once you have an Amazon S3 object registered in Collibra Catalog, it will be profiled and you can link it with policies, add data quality measures, and provide business context. All of this extra information makes it easy for the business user to identify whether the data is useful to their particular role. If so, they can add it to their shopping cart in Collibra and follow the checkout procedure. In the background, Collibra checks the applicable policies and requests the input from the necessary people to get the access approved.

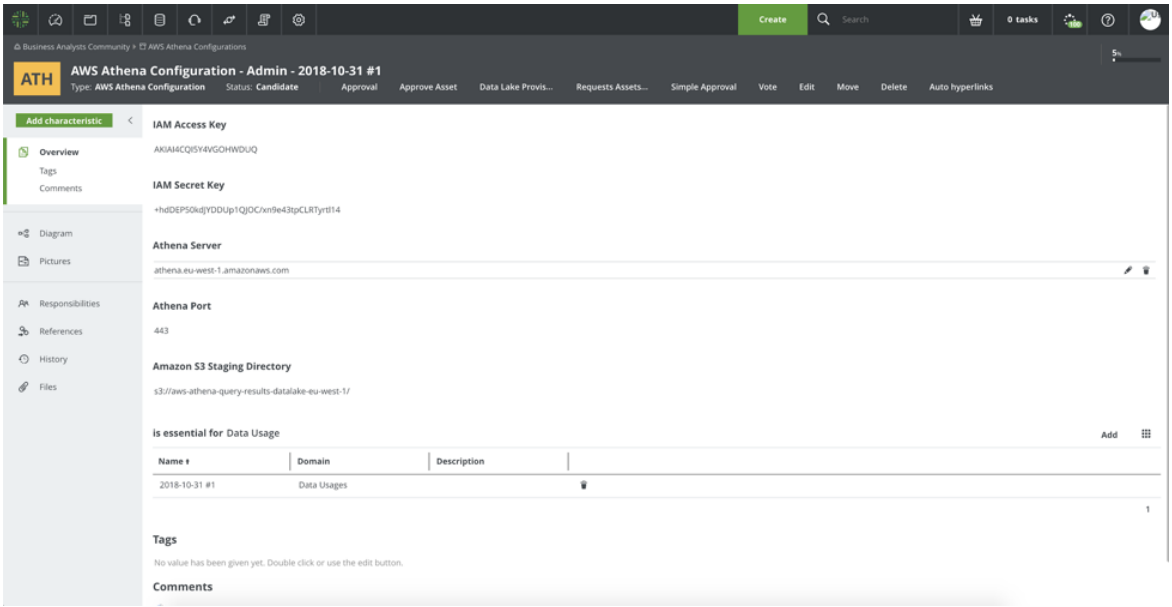
Once completed, Collibra gets an access key for this specific request so the requester can access the data they need. Once the approved time period for this access has expired, Collibra removes access using this access key.



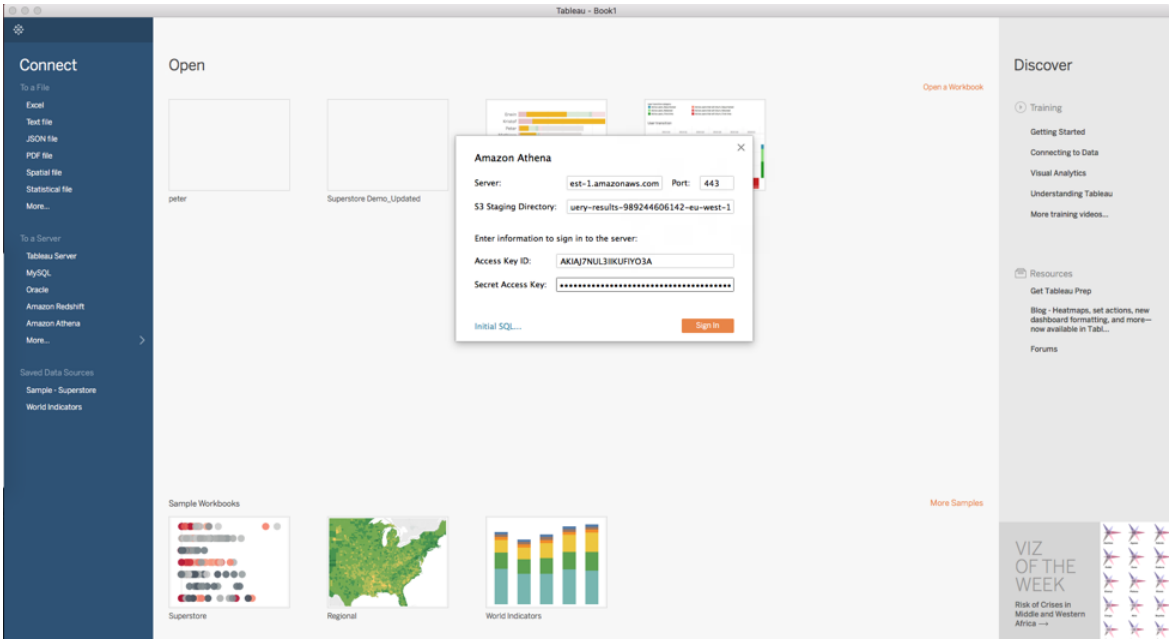
Request access to S3 data



Access approved, Collibra provides access information

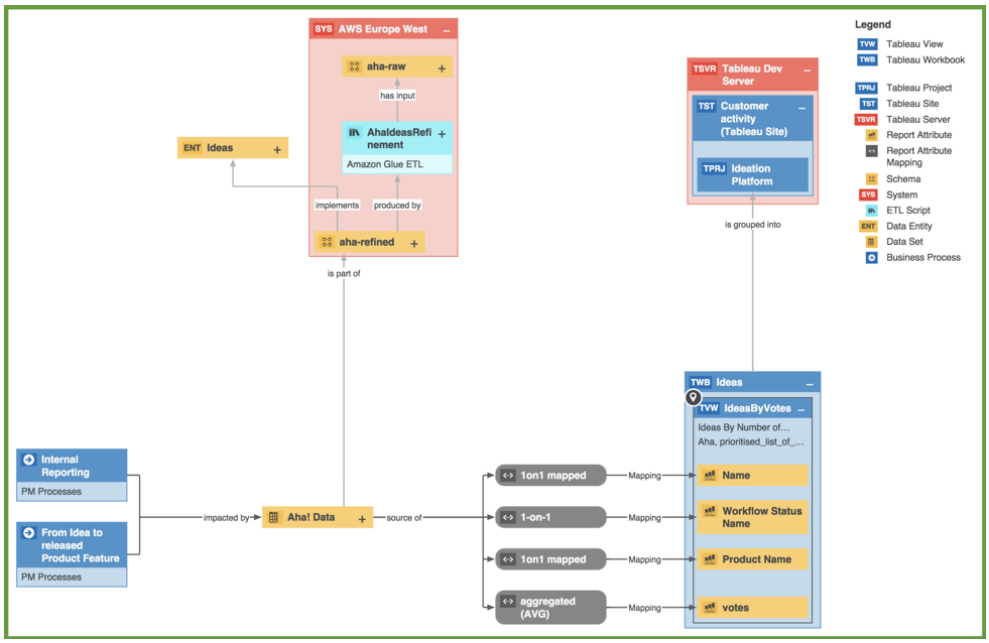


Additionally, Collibra visualizes the source to usage lineage of any S3 object. Therefore, if a user accessed the content of a file on S3 to create a report in Tableau in the past, for example, this will be visible in the lineage provided by Collibra. Our integration with analytics tools allows Collibra to generate this link automatically without the intervention of a data steward.

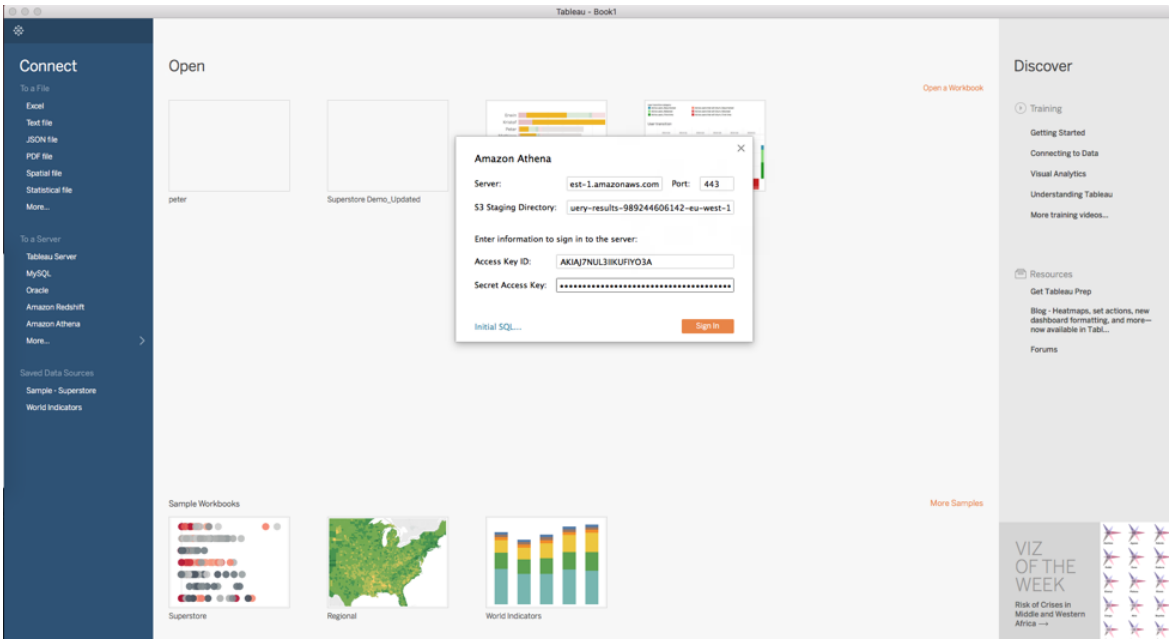


Above provided credentials used through Tableau

The partnership with Amazon delivers the most benefits to users that use AWS Glue as their preferred ETL tool. Collibra reads the Glue ETL scripts to visualize the ETL lineage from Glue in their general lineage view. So, not only will you be able to view the relationship between a report and its source data, but you will also be able to see the transformations that occurred throughout the stages of the data life cycle.



Wide lineage of Tableau report using S3 data



Lineage of Amazon Glue ETL process

Conclusion

So, how does the Collibra Catalog partnership with Amazon change the day-to-day of your user? It offers an entry point to the data landscape that allows them to find, understand, and trust their data. It provides a unified and organized view of the scattered data landscape, encouraging the user to immediately refer to the Collibra Catalog for any data-related matter. And because the catalog offers business context, users can quickly understand the data they are using. They will know how to use the data appropriately because it is linked to the proper applicable policies, and the stewardship will give users a sense of the human aspect of the data. Thanks to data shopping, they don't have to wait to get access to the data they need.

Find available data
Quickly find the right actionable data (reports, data sets, systems, …) in your enterprise data catalog

Understand your data
Easily understand your data through discovery, profiling, automated classification and data samples

Identify trusted data
Have visual access to the quality and policies related to your data

Collaborate around your data
Through crowdsourcing with your peers and data experts to become more effective as a data citizen. Tap into the tribal knowledge of people through tags, comments, ratings and repeatable workflows

Activate your data
Automate the requesting and granting of access to cloud data leveraging services like Amazon Athena and AWS Identity and Access Management

collibra®

collibra.com

US +1 646 839 3042
info@collibra.com

Follow Us:
twitter.com/collibra