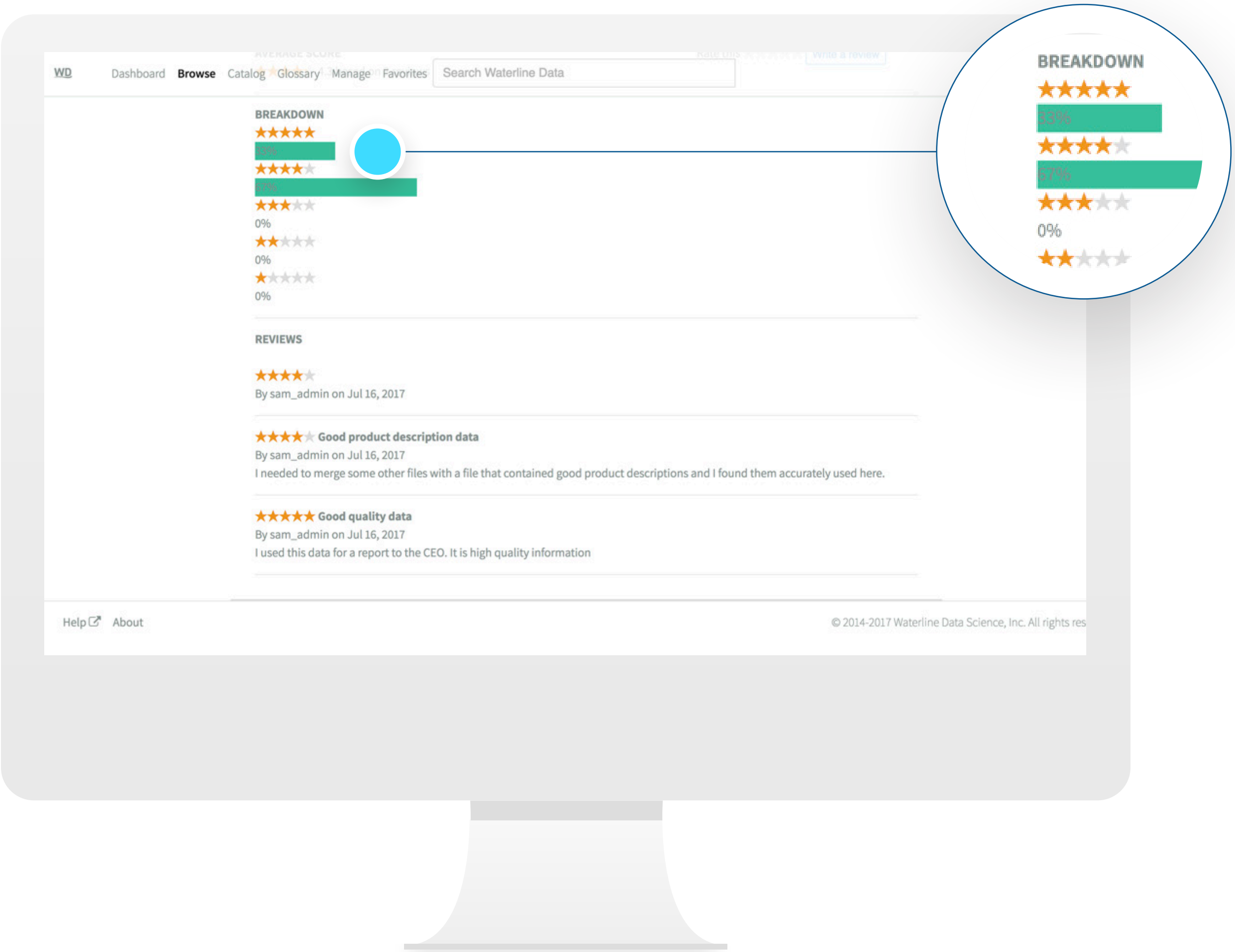
WHITE PAPER

TOP 10 CAPABILITIES TO LOOK FOR IN A DATA CATALOG

Waterline Data



BREAKDOWN

★★★★★
33%
★★★★☆
67%
★★★☆☆
0%
★★☆☆☆
0%
★☆☆☆☆
0%

REVIEWS

★★★★☆
By sam_admin on Jul 16, 2017

★★★★☆ **Good product description data**
By sam_admin on Jul 16, 2017
I needed to merge some other files with a file that contained good product descriptions and I found them accurately used here.

★★★★★ **Good quality data**
By sam_admin on Jul 16, 2017
I used this data for a report to the CEO. It is high quality information

© 2014-2017 Waterline Data Science, Inc. All rights res

The #1 Challenge in Successfully Deploying a Data Catalog

The data cataloging space is relatively new. As a result, many organizations don't know what to look for in a data catalog. Below is our list of items that we believe are critical capabilities when implementing a data catalog across your enterprise data fabric.

1.

Automated population of the catalog

The hardest problem in making a data catalog valuable is getting the catalog populated with information. And the specific information people are looking for are the tags that connect business terms to the actual attributes scattered around the organization. For the majority of businesses, there is simply too much data in their environment to be able to realistically tag the actual attributes with business terms by hand. Even crowdsourcing is not enough. Plus, manual tagging will miss dark data that hasn't been touched recently. Data tagging needs to be done by AI and machine learning as data is profiled and objective metadata about the quality of the data is added to the catalog.

data tagging needs to be done by AI and machine learning

2.

Crowdsourced curation of tags with machine learning feedback

Computerized tagging by itself is not enough. Even with the world's best tagging algorithms, we still think you will need human review to catch machine errors. Additionally, when a data steward accepts or rejects a tag, the machine learning should use that information to improve future automated tagging of data. This feedback loop is critical to improve the accuracy of the automation.

3.

Crowdsourced ratings and reviews

Context counts. Users may like some data sets more or less than others depending on the context of their job and of the data. Allowing users to rate the data with a 5 star system, as well as add written comments, provides subjective information about data sets to augment more objective profiling data added during the automated tagging process. This is kind of like a Yelp for your data!.

4.

Ability to ensure tagging and metadata freshness

It is one thing to profile and tag data once. But new data is coming into your environment all the time. You need to be able to incrementally evaluate and tag new data as it arrives and keeps your metadata fresh. This is especially important with new tag-based security paradigms, where security policies are based on metadata tags and tags can be discovered automatically by the catalog. For example, each new data set should be automatically scanned for sensitive data and tagged appropriately, so tag-based security policies can be applied.

5.

Enterprise scalability and scope

Only catalogs natively developed on big data technology like Spark and Solr can scale to process data in the entire enterprise. For example, one Waterline customer started their first effort with the Waterline data catalog by covering over 4 billion rows of data. Because Waterline takes advantage of existing Hadoop, Spark or cloud infrastructure, it is able to scale out to deal with this large volume of data, profiling and completing initial automated tagging in about 2 hours on 10 nodes. In addition to scaling to support large data volumes, a data catalog must support a wide variety of data sources in the enterprise whether on premise, in the cloud or hybrid. A data catalog needs to document all data sets in any format, be it relational, semi-structured or unstructured.

initial automated tagging in about 2 hours on 10 nodes

6.

Integration with native security infrastructure

Data catalogs cannot impose a new user management and authorization system across all the data assets. While the ability to provide a metadata-only view of the data assets is required, the ability to protect data access by respecting native authorization policies and authentication process is just as critical.

7.

Open APIs for integration with a wide variety of tools

Ranging from existing business glossaries to data wrangling tools to business intelligence tools, data catalogs have to be able to integrate with a wide variety of applications. In addition, the APIs need to support the integration of the catalog with your own applications. Many of our customers integrate the data catalog profiling and tagging capabilities as part of an automated data pipeline. APIs are critical for making this possible.

8.

Scalable search

As more users get access to the catalog, and the catalog gets larger with more tags and more profiling metadata, the ability to search the catalog will need to scale as well. That is why Waterline uses Solr for our search engine. It is a very well-known search infrastructure that has been proven to scale.

9.

Data catalog as a platform

A data catalog isn't just another piece of middleware. It is the underlying platform for a wave of new metadata-based applications. New kinds of data governance, data rationalization and consent management applications will be built on top of the data catalog in the near future. If your data catalog vendor doesn't have a vision for these kinds of apps, then you are talking to a follower, not a leader in the space.

10.

Data lineage

Being able to see at a glance where a data set came from and how it is used to generate other data sets, combined with the ability to quickly review those data sets, is key to understanding the data set and trusting it to do the job. Unfortunately, not all lineage can be imported from the tools that generate new data sets. While some can be imported from ETL tools or Hadoop systems like Apache Atlas and Cloudera Navigator, there are always gaps in lineage chains that need to be filled. The data catalog should assist in filling these gaps by automatically discovering and suggesting missing lineage between data sets. By filling in the gaps, the catalog is able to provide search based on which sources a particular data set was derived from (e.g., ability to find all customer data sets that were derived from salesforce.com regardless of how long those lineage chains are and how many data sources they span)

the data catalog should assist in filling these gaps by automatically discovering and suggesting missing lineage between data sets

Conclusion

As noted in the opening paragraph, data catalogs are a relatively new space. As time progresses, we intend to update this list to keep it current. Regardless, as of July 2017, we are confident that the information above represents the state of the art in data catalog capabilities.

Waterline Data

waterlinedata.com

Sales

(800) 123-4567
sales@waterlinedata.com

Technical Support

Visit the Support Center
help@waterlinedata.com

Corporate Headquarters

201 San Antonio Circle Suite 260
Mountain View CA 94040
(650) 946-2104