Waterline Data

Business Professionals:

Spend less time searching and more time working. Waterline makes it easy to quickly search for and find the high quality data you need to do your job.
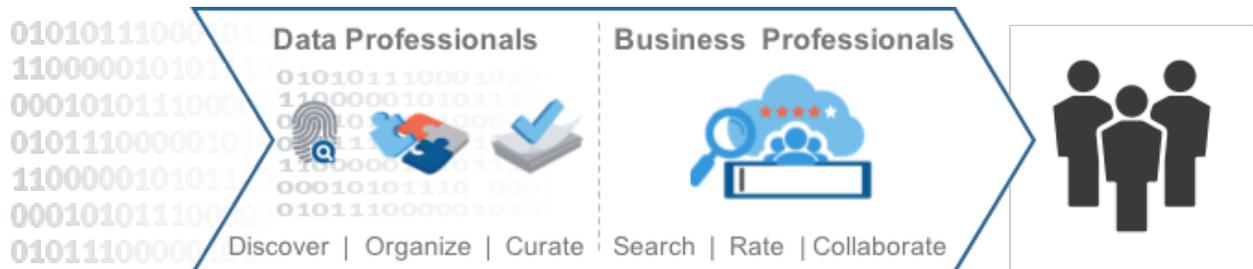
Data Professionals:

Waterline reduces the time you have to spend manually discovering, tagging, cataloging and curating data so you can get new data to business users more quickly.

Smart Data Catalog

There is so much data distributed across organizations that data and business professionals don't know what data is available or valuable. When it's time to create a new report or a dashboard for the CxO, or you are trying to respond to a request from a government regulator, or you are just trying to figure out if you can eliminate some of the excess data in your organization, the first step is usually a mad scramble to understand your existing data environment.

The Waterline Smart Data Catalog discovers and raises trusted data above the waterline so you have the data you need to effectively run your organization. We automate the discovery, matching and tagging process and ensure the catalog is always up to date by incrementally scanning the data itself and not just historical SQL logs.

- *Discover:* Waterline Data Fingerprinting™ automatically & incrementally "fingerprints" your data establishing a unique signature for each attribute. It analyzes data values, profiles the data and uses that information to do a preliminary organization of your data.

- *Organize:* Waterline automatically matches the data fingerprints to glossary terms. Unmatched terms are then matched using crowdsourcing. Once a field is tagged by an analyst, the fingerprint of that field is matched against fingerprints of other fields and the same tag is automatically propagated.

- *Curate:* The automated matches are reviewed by data analysts or data stewards for approval. The system learns from this process and tunes the automated matching algorithm. Sensitive and private data tags are passed directly to Apache Ranger and Cloudera Sentry to enable tag based access control.

- *Search, Rate and Collaborate:* Business professionals can search for data and data redundancy using the Waterline Smart Data Catalog search UI. They can also access the catalog through integrated third party data wrangling or business intelligence applications, all in a controlled and governed manner.

Key Challenges and Capabilities

For Business Professionals

CHALLENGES:

You waste too much time looking for trusted data you need to do your job.
You must ask data professionals (IT) every time you need access to new sources of data

CAPABILITIES:

Easy to use web search interface designed specifically for the business user to search a catalog of trusted, curated data and data lineage

Allows you to search from existing data wrangling and visualization tools you already use integrating with direct interfaces and REST APIs

Add annotations and view the comments of other users to enable collaboration and establish trusted data sources

Automates discovery and tagging of new data with Waterline Data Fingerprinting™, which saves time and ensures your catalog is up to date

For Data Professionals

CHALLENGES:

You spend too much time manually finding and tagging data for use by business colleagues
You are tired of having to prep and secure data to ensure proper governance

CAPABILITIES:

Reduce the time you spend manually discovering, tagging, cataloging and curating data so you can get new data to business users more quickly.

- Waterline Data Fingerprinting™ provides sophisticated data profiling algorithms that automate data tagging and data lineage discovery across all data sources

- Machine learning and crowdsourcing techniques match data fingerprints against glossary terms including sensitive data terms

Waterline simplifies data governance by delivering an automated and repeatable process to identify and ensure proper use of sensitive data:

- Data and data lineage are incrementally fingerprinted and sensitive data is automatically discovered so your data catalog is always up to date

- Replicated data is also discovered so excess data can be identified and eliminated reducing storage and management costs.

Execution Environments

Waterline Smart Data Catalog can run within the following environments:

Spark, Cloudera CDH, Hortonworks HDP, Amazon EMR, MapR, Infosys IIP, Microsoft Azure

Data Stores

Waterline Smart Data Catalog can connect to the following data sources:

Current: HDFS, HIVE, Teradata, Oracle, MySQL, MSSQL, Redshift, S3, MS ADLS, MS Blob store

Future: Any JDBC connected relational data store can be added. Contact us!