Data Governance for the Real-Time Data Lake

Data isn’t immune from data governance policies just because it’s in Hadoop. In fact, the compliance risk is greater in Hadoop because of the access to so much data as well as the risk of exposing sensitive data when different data sets are combined. Hadoop requires a data governance foundation to be able to identify and protect sensitive data, audit the data lineage, assess data quality, maintain consistent metadata, and identify any compliance issues.

MapR Technologies and Waterline Data deliver a joint solution that provides data governance capabilities on big data. The solution enables:

- **Automated data inventory.** Automatically profiles data, catalogs all the files, infers the meaning of and tags fields automatically, and detects schema changes.
- **Data quality.** Automatically generates field-level data quality assessment.
- **Metadata.** Automatically discovers extended file and field-level technical, business, and compliance metadata, and discovers data lineage and audit history.
- **Business glossary.** Crowdsources creation of ontology from business analysts working with files, and provides facilities for data stewards to edit and manage the content.
Data Governance for the Real-Time Data Lake

MapR and Waterline Data deliver a solution to enable data governance in Hadoop that meets the high availability, reliability, and real-time expectations of enterprises. The solution works at scale by automating data discovery across the data lake, and provides the foundational capabilities to profile data, assess data quality, discover and manage metadata, including an automated data inventory, automated discovery of business and compliance metadata, and self-service to find and provision the best and most trusted data for use in the target end-user tool.

Waterline Data automates the creation and management of an inventory of trusted data assets at the field level, empowering data architects to provide the right data the business needs through secure self-service.

Waterline Data assesses data quality and enables data engineers and data scientists to inspect and annotate the data, review data lineage, as well as perform impact analysis due to schema changes, before wrangling and analyzing the data in order to avoid significant downstream mistakes and risk.

Waterline Data discovers technical, business, and compliance metadata, automatically provides an integrated business glossary, and enables data stewards to manage the ontology crowdsourced from data science and big data analytics projects.

Waterline Data was founded in 2013, and is backed by Menlo Ventures and Sigma West. The inspiration for the name “Waterline” came from the metaphor of the data lake where the data is hidden below the waterline. The mission of Waterline Data is to help data engineers and data scientists find the best suited and most trusted data without coding and manual exploration—in other words, they should be able to “Hadoop above the waterline.” Waterline Data was developed to leverage the power and scalability of Hadoop to automate the inventory of data assets in the data lake and enable self-service with governance, so business users can find and understand the data in a secure and compliant way.

MapR delivers on the promise of Hadoop with a proven, enterprise-grade platform that supports a broad set of mission-critical and real-time production uses. MapR brings unprecedented dependability, ease-of-use and world-record speed to Hadoop, NoSQL, database and streaming applications in one unified distribution for Hadoop. MapR is used by more than 700 customers across financial services, government, healthcare, manufacturing, media, retail and telecommunications as well as by leading Global 2000 and Web 2.0 companies. Investors include Google Capital, Lightspeed Venture Partners, Mayfield Fund, NEA, Qualcomm Ventures and Redpoint Ventures.