Data Governance Best Practices including Automated Metadata Generation

Thomas Ward, AI Project Leader, IBM Global Chief Data Office, IBM

Category: Data Tagging

An effective and efficient data catalog supports multiple functionalities:

- Curation tools for annotating and cataloging incoming data properly
- Enforcement governance policies for data quality and standards compliance
- Extensions and enrichment to the business glossary
- Search and exploration of the data in the catalog

As data catalogs grow in size and complexity, automated AI solutions are critically required to scale these functions. IBM Watson® Knowledge Catalog (WKC), powered by IBM Cloud Pak™ for Data, is a data catalog that tightly integrates with an enterprise data governance platform. Data catalogs can help data citizens easily find, prepare, understand and use the data they need.

Watson Knowledge Catalog(WKC) helps business users quickly discover, curate, categorize and share data assets, data sets, analytical models and their relationships with other members of your organization. It serves as a single source of truth for data engineers, data stewards, data scientists and business analysts to gain self-service access to data they can trust. With data governance, data quality and active policy management, WKC helps your organization protect and govern sensitive data, trace data lineage and manage data lakes.

Automated Metadata Generation (AMG) automates the process of discovering, organizing, and curating data using Deep Learning technologies. AMG offers suggested metadata labels by looking for patterns in field-level data with technical metadata. AMG enhances speed to access and understand data within WKC. AMG has delivered to IBM's Chief Data Office running on a Cgnitive Enterprise Data Platform (CEDP); a 90% reduction in cycle time for meta data analysis, resulting in $27M in productivity savings over the past two years. AMG has dramatically enhanced data quality with regulatory and governance checks.

Metadata is just as important as data. It is the underpinning necessary in today's data era to derive meaningful business insights. Every enterprise struggles with the problem of labeling massive amounts of data. It's usually a labor-intensive manual process completed by several Subject Matter Experts (SMEs) that can take weeks. With the explosion of data from several technologies, its critical to have metadata definitions.

Classifying data, such as sensitive data is a required step to meet regulatory compliance such as GDPR or Government Owned Entities (GOE). This enables the right course of action in the handling of this data. To address the scale and speed necessary in data labeling, Artificial Intelligence is a key foundational technology.

This session will describe the main characteristics, components and approaches to building and maintaining a catalog integrated with the data lake. This session will specifically cover:

- The components, characteristics of a catalog
- The approaches to building and maintaining the business terms in a catalog
- How the catalog is used to govern the data lake assets
- How the catalog is used to support self service business insights and other user activities
- The role of AI/ML in delivering further levels of automation and efficiency to this catalog
- How the catalog can leverage graph technology and ontologies.