O'REILLY®

Compliments of ZALONI THE DATA LAKE COMPANY

Understanding Metadata

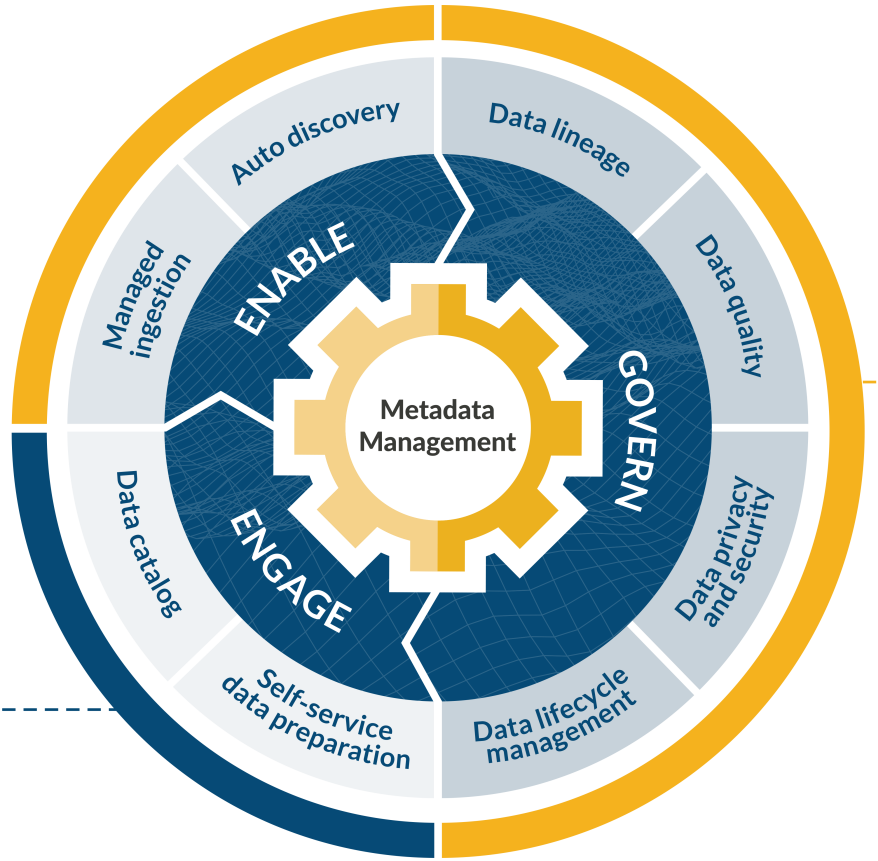
Create the Foundation for a Scalable Data Architecture

Federico Castanedo
& Scott Gidley

ZALONI
THE DATA LAKE COMPANY

Data Lake 360° – A holistic approach to managing big data

BEDROCK
DATA LAKE MANAGEMENT
& GOVERNANCE PLATFORM



MICA
SELF-SERVICE
DATA PLATFORM

To learn more:

Call us: 1 919.323.4050 | E-mail: info@zaloni.com | Visit: www.zaloni.com

Understanding Metadata

Create the Foundation for a Scalable Data Architecture

Federico Castanedo and Scott Gidley

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Understanding Metadata
by Federico Castanedo and Scott Gidley

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://www.oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Shannon Cutt
Production Editor: Colleen Lobner
Copyeditor: Charles Roumeliotis
Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

February 2017: First Edition

Revision History for the First Edition
2017-02-15: First Release

978-1-491-97486-5

[LSI]

Table of Contents

CHAPTER 1

Understanding Metadata: Create the Foundation for a Scalable Data Architecture

Key Challenges of Building Next-Generation Data Architectures

Today's technology and software advances allow us to process and analyze huge amounts of data. While it's clear that big data is a hot topic, and organizations are investing a lot of money around it, it's important to note that in addition to considering scale, we also need to take into account the *variety* of the types of data being analyzed. Data *variety* means that datasets can be stored in many formats and storage systems, each of which have their own characteristics. Taking data variety into account is a difficult task, but provides the benefit of having a *360-degree approach*—enabling a full view of your customers, providers, and operations. To enable this 360-degree approach, we need to implement next-generation data architectures. In doing so, the main question becomes: *how do you create an agile data platform that takes into account data variety and scalability of future data?*

The answer for today's forward-looking organizations increasingly relies on a data lake. A *data lake* is a single repository that manages transactional databases, operational stores, and data generated outside of the transactional enterprise systems, all in a common repository. The data lake supports data from different sources like files,

clickstreams, IoT sensor data, social network data, and SaaS application data.

A core tenet of the data lake is the storage of raw, unaltered data; this enables flexibility in analysis and exploration of data, and also allows queries and algorithms to evolve based on both historical and current data, instead of a single point-in-time snapshot. A data lake also provides benefits by avoiding information silos and centralizing the data into one common repository. This repository will most likely be distributed across many physical machines, but will provide end users transparent access and a unified view of the underlying distributed storage. Moreover, data is not only distributed but also *replicated*, so access, redundancy, and availability can be ensured.

A data lake stores all types of data, both structured and unstructured, and provides *democratized access* via a single unified view across the enterprise. In this approach you can support many different data sources and data types in a single platform. A data lake strengthens an organization's existing IT infrastructure, integrating with legacy applications, enhancing (or even replacing) an enterprise data warehouse (EDW) environment, and providing support for new applications that can take advantage of the increasing data variety and data volumes experienced today.

Being able to store data from different input types is an important feature of a data lake, since this allows your data sources to continue to evolve without discarding potentially valuable metadata or raw attributes. A breadth of different analytical techniques can also be used to execute over the same input data, avoiding limitations that arise from processing data only after it has been aggregated or transformed. The creation of this unified repository that can be queried with different algorithms, including SQL alternatives outside the scope of traditional EDW environments, is the hallmark of a data lake and a fundamental piece of any big data strategy.

To realize the maximum value of a data lake, it must provide (1) the ability to ensure data quality and reliability, that is, ensure the data lake appropriately reflects your business, and (2) easy access, making it faster for users to identify which data they want to use. To govern the data lake, it's critical to have processes in place to cleanse, secure, and operationalize the data. These concepts of data governance and data management are explored later in this report.

Building a data lake is not a simple process, and it is necessary to decide *which* data to ingest, and how to organize and catalog it. Although it is not an automatic process, there are tools and products to simplify the creation and management of a modern data lake architecture at enterprise scale. These tools allow ingestion of different types of data—including streaming, structured, and unstructured; they also allow application and cataloging of metadata to provide a better understanding of the data you already ingested or plan to ingest. All of this allows you to create the foundation for an *agile data lake platform*.

For more information about building data lakes, download the free O'Reilly report *Architecting Data Lakes*.

What Is Metadata and Why Is It Critical in Today's Data Environment?

Modern data architectures promise the ability to enable access to more and different types of data to an increasing number of data consumers within an organization. Without proper governance, enabled by a strong foundation of *metadata*, these architectures often show initial promise, but ultimately fail to deliver.

Let's take *logistics distribution* as an analogy to explain *metadata,* and why it's critical in managing the data in today's business environment. When you are shipping one package to an international destination, you want to know where in the route the package is located in case something happens with the package delivery. Logistic companies keep manifests to track the movement of packages and the successful delivery of packages along the shipping process.

Metadata provides this *same type of visibility* into today's data rich environment. Data is moving in and out of companies, as well as within companies. Tracking data changes and detecting any process that causes problems when you are doing data analysis is hard if you don't have information about the data and the data movement process. Today, even the change of a single column in a source table can impact hundreds of reports that use that data—making it extremely important to know *beforehand* which columns will be affected.

Metadata provides information about each dataset, like size, the schema of a database, format, last modified time, access control lists, usage, etc. The use of metadata enables the management of a scala-

ble data lake platform and architecture, as well as *data governance*. Metadata is commonly stored in a central catalog to provide users with information on the available datasets.

Metadata can be classified into three groups:

- *Technical metadata* captures the form and structure of each dataset, such as the size and structure of the schema or type of data (text, images, JSON, Avro, etc.). The structure of the schema includes the names of fields, their data types, their lengths, whether they can be empty, and so on. Structure is commonly provided by a relational database or the heading in a spreadsheet, but may also be added during ingestion and data preparation. There are some basic technical metadata that can be obtained directly from the datasets (i.e., size), but other metadata types are derived.
- *Operational metadata* captures the lineage, quality, profile, and provenance (e.g., when did the data elements arrive, where are they located, where did they arrive from, what is the quality of the data, etc.). It may also contain how many records were rejected during data preparation or a job run, and the success or failure of that run itself. Operational metadata also identifies how often the data may be updated or refreshed.
- *Business metadata* captures what the data means to the end user to make data fields easier to find and understand, for example, business names, descriptions, tags, quality, and masking rules. These tie into the business attributes definition so that everyone is consistently interpreting the same data by a set of rules and concepts that is defined by the business users. A business glossary is a central location that provides a business description for each data element through the use of *metadata information.*

  Metadata information can be obtained in different ways. Sometimes it is encoded within the datasets, other times it can be inferred by reading the content of the datasets; or the information can be spread across log files that are written by the processes that access these datasets.

In all cases, metadata is a key element in the management of the data lake, and is the foundation that allows for the following data lake characteristics and capabilities to be achieved:

- *Data visibility* is provided by using metadata management to keep track of what data is in the data lake, along with source, format, and lineage. This can also include a time series view, where you can see what actions were assigned or performed and see exclusions and inclusions. This is very useful if you want to do an impact analysis, which may be required as you're doing change management or creating an agile data platform.
- *Data reliability* gives you confidence that your analytics are always running on the right data, with the right quality, which may also include analysis of the metadata. A good practice is to use a combination of top-down and bottom-up approaches. In the top-down approach, a set of rules defined by business users, data stewards, or a center of excellence is applied, and these rules are stored as metadata. On the other hand, in the bottom-up approach, data consumers can further qualify or modify the data or rate the data in terms of its usability, freshness, etc. Collaboration capabilities in a data platform have become a common way to leverage the "wisdom of crowds" to determine the reliability of data for a specific use case.
- *Data profiling* allows users to obtain information about specific datasets and to get a sense for the format and content of the data. It enables data scientists and business analysts a quick way to determine if they want to use the data. The goal of data profiling is providing a view for end users that helps them understand the content of the dataset, the context in which it can be used in production, and any anomalies or issues that might require remediation or prohibit use of the data for further consumption. In an agile data platform, data profiling should scale to meet any data volume, and be available as an automated process on data ingest or as an ad hoc process available to data scientists, business analysts, or data stewards who may apply subject matter expertise to the profiling results.
- *Data lifecycle/age*: You are likely to have different aging requirements for the data in your data lake, and these can be defined by using operational metadata. Retention schemes can be based on global rules or specific business use cases, but are always aimed

at translating the value of data at any given point into an appropriate storage and access policy. This maximizes the available storage and gives priority to the most critical or high-usage data. Early implementations of data lakes have often overlooked data lifecycle as the low cost of storage and the distributed nature of the data made this a lower priority. As these implementations mature, organizations are realizing that *managing* the data lifecycle is critical for maintaining an effective and IT compliant data lake.

- *Data security and privacy*: Metadata allows access control and data masking (e.g., for personally identifiable information (PII)), and ensures compliance with industry and other regulations. Since it is possible to define what datasets are sensitive, you can protect the data, encrypt columns with personal information, or give access to the right users based on metadata. Annotating datasets with security metadata also simplifies audit processes, and helps to expose any weaknesses or vulnerabilities in existing security policies. Identification of private or sensitive data can be determined by integrating the data lake metadata with enterprise data governance or business glossary solutions, introspecting the data upon ingest to look for common patterns (SSN, industry codes, etc.), or utilizing the data profiling or data discovery process.
- *Democratized access to useful data*: Metadata allows you to create a system to extend end-user accessibility and self-service (to those with permissions) to get more value from the data. With an extensive metadata strategy in place, you can provide a robust catalog to end users, from which it's possible to search and find data on any number of facets or criteria. For example, users can easily find customer data from a Teradata warehouse that contains PII data, without having to know specific table names or the layout of the data lake.
- *Data lineage and change data capture*: In current data production pipelines, most companies focus only on the metadata of the input and output data, enabling the previous characteristics. However, it is common to have several processes between the input and the output datasets, and these processes are not always managed using metadata, and therefore do not always capture data change or lineage. In any data analysis or machine learning process, the results are always obtained from the com-

bination of running specific algorithms over particular datasets, so it becomes extremely important to have metadata information about the intermediate processes, in order to enhance or improve it over time.

Data lakes must be architected properly to leverage metadata and integrate with existing metadata tools, otherwise it will create a hole in organizations' data governance process because how data is used, transformed, and related outside the data lake can be lost. An incorrect metadata architecture can often prevent data lakes making the transition from an analytical sandbox to an enterprise data platform.

Ultimately, most of the time spent in data analysis is in preparing and cleaning the data, and metadata helps to reduce the time to insight by providing easy access to discovering what data is available, and maintaining a full data tracking map (data lineage).

A Modern Data Architecture—What It Looks Like

Unlike a traditional data architecture driven by an extract, transform, load (ETL) process that loads data into a data warehouse, and then creates a rationalized data model to serve various reporting and analytic needs, data lake architectures look very different. Data lakes are often organized into *zones* that serve specific functions.

The data lake architecture begins with the ingestion of data into a staging area. From the staging area it is common to create new/different transformed datasets that either feed net-new applications running directly on the data lake, or if desired, feed these transformations into existing EDW platforms.

Secondly, as part of the data lake, you need a framework for capturing metadata so that you can later leverage it for various use case functionalities discussed in the previous section. The big data management platform of this modern data architecture can provide that framework.

The key is being able to automate the capture of metadata on arrival, as you're doing transformations, and tying it to specific definitions like the enterprise business glossary.

Managing a modern data architecture also requires attention to data lifecycle issues like *expiration* and *decommissions* of data and to ensure access to data within specific time constraints.

In Figure 1-1, we'll take a closer look at an example of how a modern data architecture might look; the example in Figure 1-1 comes from Zaloni.

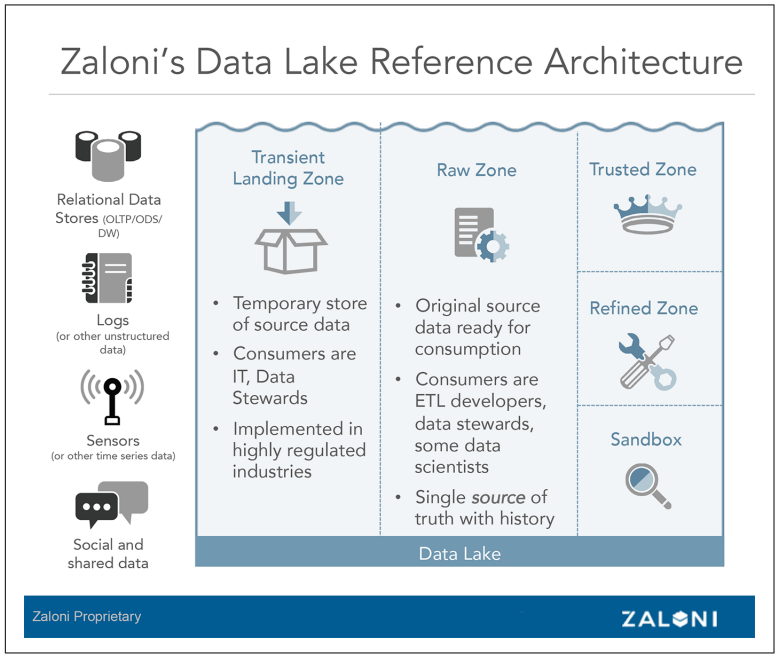


Figure 1-1. A sample data lake architecture from Zaloni

To the left, you have different data sources that can be ingested into the data lake. They may be coming through an ingestion mechanism in various different formats whether they are file structures, database extracts, the output of EDW systems, streaming data, or cloud-based REST APIs.

As data comes into the lake (blue center section), it can land in a Transient Zone (a.k.a. staging area) before being made consumable to additional users, or drop directly into the Raw Zone. Typically, companies in regulated industries prefer to have a transient loading zone, while others skip this zone/stage.

In the Raw Zone the data is kept in its original form, but it may have sensitive data masked or tokenized, to ensure compliance with secu-

rity and privacy rules. Metadata discovery upon ingest can often identify PII or sensitive data for masking. Next, after applying metadata you have the flexibility to create other useful zones:

- *Refined Zone*: Based on the metadata and the structure of the data, you may want to take some of the raw datasets and transform them into refined datasets that you may need for various use cases. You also can define some new structures for your common data models and do some data cleansing or validation using metadata.
- *Trusted Zone*: If needed, you could create some master datasets and store them in what is called "the Trusted Data Zone" area of the data lake. These master data sets may include frequently accessed reference data libraries, allowable lists of values, or product or state codes. These datasets are often combined with refined datasets to create analytic data sets that are available for consumption.
- *Sandbox*: An area for your data scientists or your business analysts to play with the data and again, leverage metadata to more quickly know how fresh the datasets are, assess the quality of the data, etc., in order to build more efficient analytical models on top of the data lake.

Finally, on the righthand side of the sample architecture, you have the Consumption Zone. This zone provides access to the widest range of users within an organization.

Data Lake Management Solutions

For organizations considering a data lake, there are big data tools, data management platforms, and industry-specific solutions available to help meet overall data governance and data management requirements. Organizations that are early adopters or heavily IT-driven may consider building a data lake by stitching together the plethora of tooling available in the big data ecosystem. This approach allows for maximum flexibility, but incurs higher maintenance costs as the use cases and ecosystem change. Another approach is to leverage existing data management solutions that are in place, and augment them with solutions for metadata, self-service data preparation, and other areas of need. A third option is to implement an end-to-end data management platform that is built natively for the big data ecosystem.

Depending on the provider, data lake management solutions can be classified into three different groups: (1) solutions from traditional data integration/management vendors, (2) tooling from open source projects, and (3) startups providing best-of-breed technology.

Traditional Data Integration/Management Vendors

The IBM Research Accelerated Discovery Lab is a collaborative environment specifically designed to facilitate analytical research projects. This lab leverages IBM's Platform Cluster Management and includes data curation tools and data lake support. The lab provides data lakes that can ingest data from open source environments (e.g., *data.gov/*) or third-party providers, making contextual and project-specific data available. The environment includes tools to pull data from open APIs like Socrata and ckan. IBM also provides InfoSphere Information Governance Catalog, a metadata management solution that helps to manage and explore data lineage.

The main drawback of solutions from traditional data integration vendors is the integration with third-party systems; although most of them include some integration mechanism in one way or another, it may complicate the data lake process. Moreover they usually require a heavy investment in technical infrastructure and people with specific skills related to their product.

Tooling From Open Source Projects

Teradata Kylo is a sample framework for delivering data lakes in Hadoop and Spark. It includes a user interface for data ingesting and wrangling and provides metadata tracking. Kylo uses Apache NiFi for orchestrating the data pipeline. Apache NiFi is an open source project developed under the Apache ecosystem and supported by HortonWorks as DataFlow. NiFi is an integrated data logistics platform for automating the movement of data between disparate systems. It provides data buffering and provenance when moving data by using visual commands (i.e., drag and drop) and control in a web-based user interface.

Apache Atlas is another solution, currently in the incubator state. Atlas is a scalable and extensible set of core foundational governance services. It provides support for data classification, centralized auditing, search, and lineage across Hadoop components.

Oracle Enterprise Metadata Management is a solution that is part of the Fusion Middleware. It provides metadata exploration capabilities and improves data governance and standardization through metadata.

Informatica is another key player in the world of metadata management solutions with a product named Intelligent Data Lake. This solution prepares, catalogs, and shares relevant data among business users and data scientists.

Startups Providing Best-of-Breed Technology

Finally, there are some startups developing commercial products customized for data lake management, like:

- Trifacta's solution focuses on the problem of integrating and cleaning the datasets as they come into the lake. This tool essentially prepares the datasets for efficient posterior processing.
- Paxata is a data preparation platform provider that provides data integration, data quality, semantic enrichment, and governance. The solution is available as a service and can be deployed in AWS virtual private clouds or within Hadoop environments at customer sites.
- Collibra Enterprise Data Platform provides a repository and workflow-oriented data governance platform with tools for data management and stewardship.
- Talend Metadata Manager imports metadata on demand from different formats and tools, and provides visibility and control of the metadata within the organization. Talend also has other products for data integration and preparation.
- Zaloni provides Bedrock, an integrated data lake management platform that allows you to manage a modern data lake architecture, as shown in Figure 1-1. Bedrock integrates metadata from the data lake and automates metadata inventory. In Bedrock, the metadata catalog is a combination of technical, business, and operational metadata. Bedrock allows searching and browsing for metadata using any related term.

Bedrock can generate metadata based on ingestions, by importing Avro, JSON, or XML files. Data collection agents compute the metadata, and the product shows users a template to be approved with the metadata. It also automates metadata creation when you add relational databases, and can read data directly from the data lake.

With Bedrock all steps of data ingestion are defined in advance, tracked, and logged. The process is repeatable. Bedrock captures streaming data and allows you to define streams by integrating Kafka topics and flume agents.

Bedrock can be configured to automatically consume incoming files and streams, capture metadata, and register with the Hadoop ecosystem. It employs file- and record-level watermarking, making it possible to see where data moves and how it is used (*data lineage*). Input data can be enriched and transformed by implementing Spark-based transformation libraries, providing flexible transformations at scale.

One challenge that the Bedrock product addresses is metadata management in *transient clusters*. Transient clusters are configured to allow a cost-effective, scalable on-demand process, and they are turned off when no data processing is required. Since metadata information needs to be persistent, most companies decide to pay an extra cost for persistent data; one way to address this is with a data lake platform, such as Bedrock.

Zaloni also provides Mica, a self-service data preparation product on top of Bedrock that enables business users to do data exploration, preparation, and collaboration. It provides an enterprise-wide data catalog to explore and search for datasets using free-form text or multifaceted search. It also allows users to create transformations interactively, using a tabular view of the data, along with a list of transformations that can be applied to each column. Users can define a process and operationalize it in Bedrock, since Mica creates a workflow by automatically translating the UI steps into Spark code and transferring it to Bedrock.

Automating Metadata Capture

Metadata generation can be an exhausting process if it is performed by manually inspecting each data source. This process is even harder in larger companies with numerous but disparate data sources. As we mentioned before, the key is being able to automate the capture of metadata *on arrival* of data in the lake, and identify relationships with existing metadata definitions, governance policies, and business glossaries.

Sometimes metadata information is not provided in a machine-readable form, so metadata must be entered manually by the data curator, or discovered by a specific product. To be successful with a modern data architecture, it's critical to have a way to automatically register or discover metadata, and this can be done by using a metadata management or generation platform.

Since the data lake is a cornerstone of the modern data architecture, whatever metadata is captured in the data lake also needs to be fed into the enterprise metadata repository, so that you have an end-to-end view across all the data assets in the organization, including, but beyond, the data lake. An idea of what automated metadata registration could look like is shown in Figure 1-2.

Figure 1-2 shows an API that runs on a Hadoop cluster, which retrieves metadata such as origin, basic information, and timestamp and stores it in an operational metadata file. New metadata is also stored in the enterprise metadata repositories, so it will be available for different processes.

Another related step that is commonly applied in the automation phase is the encryption of personal information and the use of tokenization algorithms.

Ensuring data quality is also a relevant point to consider in any data lake strategy. *How do you ensure the quality of the data transparently to the users?*

One option is to profile the data in the ingestion phase and perform a statistical analysis that provides a quality report by using metadata. The quality can be performed at each dataset level and the information can be provided using a dashboard, by accessing the corresponding metadata.

A relevant question in the automation of metadata is *how do we handle changes in data schema*? Current solutions are just beginning to scratch the surface of what can be done here. When a change in the metadata occurs it is necessary to reload the data. But it would be very helpful to automate this process and introspect the data directly to detect schema changes in real time. So, when metadata changes, it will be possible to detect modifications by creating a new entity.

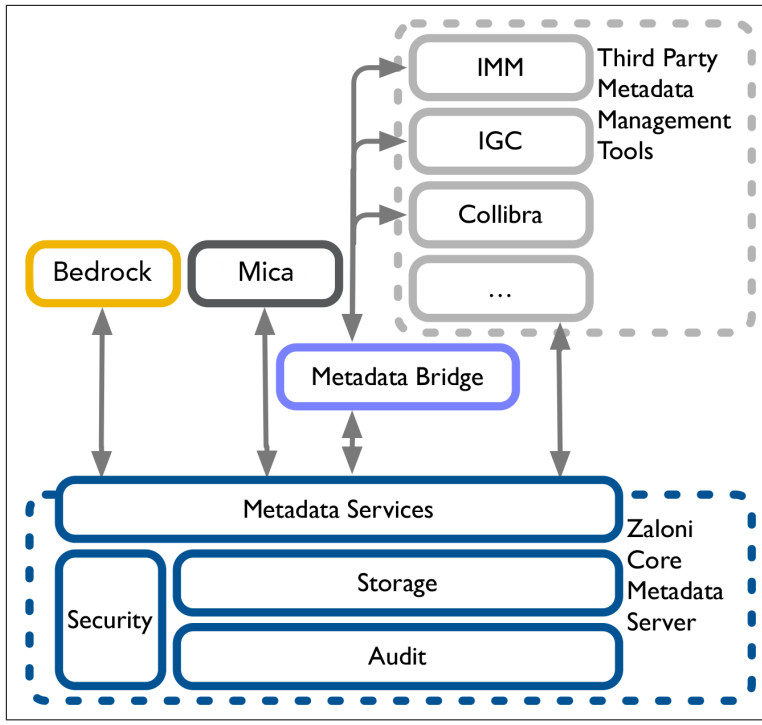


Figure 1-2. An example of automated metadata registration from Zaloni

Looking Ahead

Another level of automation is related to data processing. Vendors are looking at ways to make recommendations for the automation of data processing based on data ingestion and previous pre-processing of similar datasets. For example, at the time you ingest new data in the platform, the solution provides a suggestion to the user like "this looks like customer data and the last time you got similar data you applied these transformations, masked these fields, and set up a data lifecycle policy."

There is also an increased interest around using metadata to understand context. For example, an interesting project going on at UC Berkeley, called Ground, is looking at new ways to allow people to understand data context using open source and vendor neutral technology. The goal of Ground is to enable users to reason about what data they have, where that data is flowing to and from, who is

using the data, when the data changed, and why and how the data is changing.

Conclusion

Since most of the time spent on data analysis projects is related with identifying, cleansing, and integrating data and is magnified when data is stored across many silos, the investment in building a data lake is worthwhile. With a data lake you can significantly reduce the effort of finding datasets, the need to prepare them in order to make them ready to analyze, and the need to regularly refresh them to keep them up-to-date.

Developing next-generation data architectures is a difficult task because it is necessary to take into account the format, protocol, and standards of the input data, and the veracity and validity of the information must be ensured while security constraints and privacy are considered. Sometimes it is very difficult to build all the required phases of a data lake from scratch, and most of the time it is something that must be performed in phases. In a next-generation data architecture, the focus shifts over time from data ingestion to transformation, and then to analytics.

As more consumers across an organization want to access and utilize data for various business needs, and enterprises in regulated industries are looking for ways to enable that in a controlled fashion, metadata as an integral part of any big data strategy is starting to get the attention it deserves.

Due to the democratization of data that a data lake provides, ample value can be obtained from the way that data is used and enriched, with metadata information providing a way to share discoveries with peers and other domain experts.

But data governance in the data lake is key. Data lakes must be architected properly to leverage metadata and integrate with existing metadata tools, otherwise it will create a hole in organizations' data governance processes because how data is used, transformed, and related outside the data lake can be lost. An incorrect metadata architecture can often prevent data lakes from making the transition from an analytical sandbox to an enterprise data platform.

Building next-generation data architectures requires effective metadata management capabilities in order to operationalize the data

lake. With all of the available options now for tools, it is possible to simplify and automate common data management tasks, so you can focus your time and resources on building the insights and analytics that drive your business.

About the Authors

Federico Castanedo is the Lead Data Scientist at Vodafone Group in Spain, where he analyzes massive amounts of data using artificial intelligence techniques. Previously, he was Chief Data Scientist and cofounder at WiseAthena.com, a startup that provides business value through artificial intelligence. For more than a decade, he has been involved in projects related to data analysis in academia and industry. He has published several scientific papers about data fusion techniques, visual sensor networks, and machine learning. He holds a PhD in Artificial Intelligence from the University Carlos III of Madrid and has also been a visiting researcher at Stanford University.

Scott Gidley is Vice President of Product Management for Zaloni, where he is responsible for the strategy and roadmap of existing and future products within the Zaloni portfolio. Scott is a nearly 20-year veteran of the data management software and services market. Prior to joining Zaloni, Scott served as senior director of product management at SAS and was previously CTO and cofounder of DataFlux Corporation. Scott received his BS in Computer Science from University of Pittsburgh.